An Introduction to Logical Information Theory: A Conceptual Foundation for Shannon's Information Theory

David Ellerman

December 9, 2006

Contents

Abstract

Categorical logic has shown that logic is essentially the logic of subsets (or "subobjects"). Predicates are modelled as subsets of a universe set and a proposition formed by applying a predicate to an individual name is satisfied in the model if the named individual is in the subset representing the predicate. Since partitions are dual to subsets (i.e., epimorphisms are dual to monomorphisms), this suggests the possibility of a dual logic of partitions. We argue that the notion of a "distinction" is dual to the notion of an "element" where a distinction is modelled by an ordered pair of elements (u, u') from the universe set U. A predicate modelled by a partition π on U would apply to the name of a distinction if the pair of elements was distinguished by the partition π, i.e., if u and u' were in different blocks of π.

The next step in the development of subset logic was to develop probability theory for events starting with the assignment of the relative size to each subset-event of a finite universe. Hence

the analogous next step in a dual logic of partitions is to assign to a partition the number of distinctions made by a partition relative to the total number of ordered pairs $|U|^2$ from the finite universe. That yields a notion of "logical entropy" for partitions. This paper is an introduction to the resulting "logical information theory." The notion of a "distinction" is the conceptual atom, dual to the notion of element, which is basic building block for information theory. The logical theory directly counts the (normalized) number of distinctions in a partition while Shannon's theory gives the average number of binary partitions needed to make those distinctions. Thus the logical theory is seen as providing a new foundation or conceptual underpinning) for Shannon's theory based on the conceptual atoms of "distinctions."

1 The Idea of a Dual Logic of Partitions

1.1 Dualizing Categorical Logic?

In ordinary logic, a statement $P(a)$ is formed by a predicate $P(x)$ applying to an individual a (which could be an n-tuple in the case of relations). The predicate is modelled by a subset S of a universe set U and an individual name such as "a" would be assigned an individual $s \in U$. The statement $P(a)$ would hold in the model if $s \in S$. In short, logic is modelled as the logic of subsets of a set. Largely due to the efforts of William Lawvere, the modern treatment of logic was reformulated and vastly generalized using category theory in what is now called *categorical logic*. Subsets were generalized to subobjects or "parts" (equivalence classes of monomorphisms) so that logic has become the logic of subobjects.[1]

There is a duality between subsets of a set and partitions[2] on a set: "The dual notion (obtained by reversing the arrows) of 'part' is the notion of *partition*." [10, p. 85] In category theory, this emerges as the reverse-the-arrows duality between monomorphisms (monos), e.g., injective set functions, and epimorphisms (epis), e.g., surjective set functions, and between subobjects and quotient objects.[3] If modern logic is formulated as the logic of subsets, or more generally, subobjects or "parts", then the question naturally arises of a "cologic" that might play the analogous role for partitions and their generalizations.

Quite aside from category theory duality, it has long been noted in mathematics, e.g., in Gian-Carlo Rota's work in combinatorial theory and probability theory [2], that there is a type of duality between subsets of a set and partitions on a set. Just as subsets of a set are partially ordered by inclusion, so partitions on a set are partially ordered by refinement.[4] Moreover, both partial orderings are in fact lattices (i.e., have meets and joins) with a top element $\widehat{1}$ and a bottom element $\widehat{0}$. In the lattice of all subsets $\mathcal{P}(U)$ (the power set) of a set U, the meet and join are, of course, intersection and union while the top element is the universe U and the bottom element is the null set $\varnothing$. In the lattice of all partitions $\Pi(U)$ on a non-empty set U, there are also meet and join operations (much more on them later) while the top element is the indiscrete partition (the "blob") where all of U is one block and the bottom element is the discrete partition where each element of U is a singleton block.

The dual logic or cologic for partitions and its application to information theory ("logical information theory") involves some "rethinking" of partitions. If a partition is dual to a subset in the usual duality between epimorphisms and monomorphisms, then what is the partition-related dual to the notion of a *member* of a set? We shall argue that the dual to the notion of "element" of a subset

[1] See [10] Appendix A for a good treatment. For the vast generalization of topos theory see [12] and for the category theoretic background, see the standard reference [11].

[2] A partition π on a set U is usually defined as a mutually exclusive and jointly exhaustive set $\{B\}_{B\in\pi}$ of subsets or "blocks" $B \subseteq U$.

[3] Every equivalence relation (reflexive, symmetric, and transitive relation) on a set U determines a partition on U as the equivalence classes of the equivalence. Conversely, every partition on a set determines an equivalence relation on the set (two elements are equivalent if they are in the same block of the partition). The notions of a partition on a set and an equivalence relation on a set are equivalent.

[4] A partition f refines a partition g, written $f \preceq g$, if each block of f is contained in some block of g.

is the notion of a "distinction" of a partition. In the logic of partitions, a predicate $P(x)$ would be modelled by a partition π on U, an individual name "a" would be assigned a distinction (u, u') where $u, u' \in U$ (or n-tuples of distinctions), and the statement $P(a)$ would hold in the model if the partition made that distinction, i.e., if u and u' were in different blocks of the partition.

The logic of subsets extends to probability theory which begins with the model of events as subsets S of a finite sample space U and then assigns probabilities $\text{Prob}(S)$ to subsets (e.g., the Laplacian equiprobable distribution where $\text{Prob}(S) = |S| / |U|$). The logic of partitions similarly extends to a theory that assigns numerical values to partitions rather than subsets. That theory turns out to be a "logical" information theory where that numerical value is the logical information content or logical entropy of the partition and it is initially defined in a Laplacian manner as the number of distinctions that a partition makes normalized by the number of ordered pairs of the universe set U. This logical entropy is precisely related to Shannon's entropy measure so the development of logical information theory can be seen as providing a new conceptual foundation for information theory at the basic level of logic using the "distinctions" as the conceptual atoms.

Our purpose in this paper is to introduce logical information theory, including arguments to motivate the basic concept of a "distinction," and to develop the connection with Shannon's information theory.

1.2 The Dual 'Creation Myths' for a Set

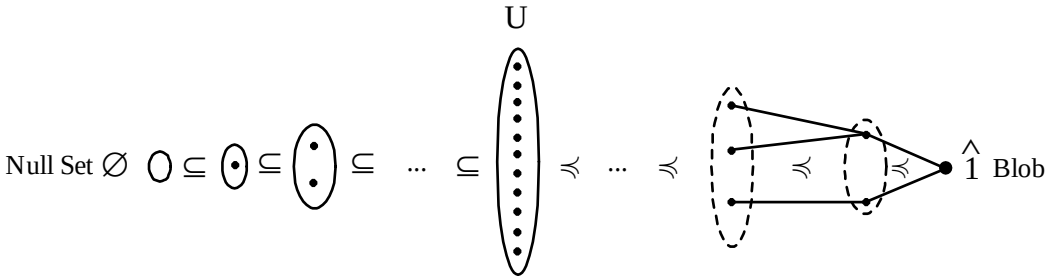
This development of the dual logic of partitions is based on a vision of how partitions on a set are dual to subsets of a set. This duality will be initially described using a pair of heuristic "creation myths." There is not one but two ways to create a set, and set theory only deals with one of them.

In the Boolean lattice of subsets of a set U, the universe set U and the null set $\varnothing$ seem to have symmetrical roles. But in the development of set theory, the notion of a universe set disappears. The null set can be considered as a subset of larger and larger sets (and thus the codomain of $\varnothing \to U$ changes) but in a certain sense the null set remains the same. In the lattice of partitions on a set U, the discrete partition $\widehat{0}$ on U and the indiscrete partition or blob $\widehat{1}$ seem to have symmetrical roles. But as larger and larger universe sets are considered, the discrete partition on the universe changes (and the domain of $U \to 1$ changes) but in a certain sense the blob remains the same. Thus the idea arises of telling two stories about sets, one story starting with the null set and the other story starting with the blob.

Each creation myth starts with an initial object. In one creation story, we start with the null set $\varnothing$, the initial object in the category **Set** of sets and functions, and then imagine objects, individuals, or elements being created over a period of time, each element having fixed characteristics to distinguish it from the other elements. At any point in time, we take a snapshot of all the elements created so far which make up a universe set U. Then the lattice $\mathcal{P}(U)$ of subsets of U consists of all the possible intermediate stages between $\varnothing$ and U.

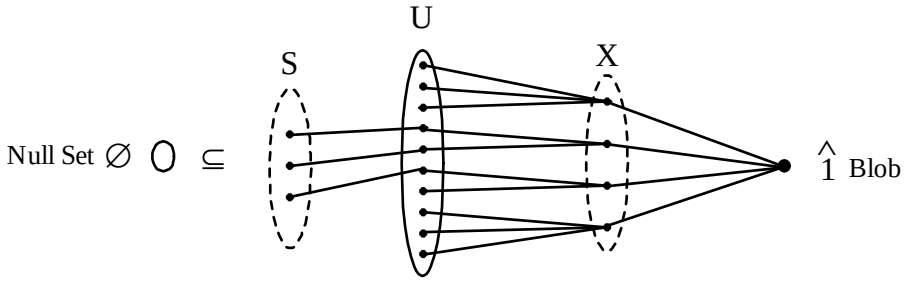
Just as the null set can be thought of as a void or empty form with no substance, so the blob can be thought of as a substance without form like a blank canvas. In the dual creation myth, we start with the initial object in the dual of the category of sets $\mathbf{Set}^{op}$ (i.e., the terminal object 1 in **Set**), which is this context is called the "blob," denoted $\widehat{1}$, and then make *distinctions* (like drawing a line down the middle of a canvas to distinguish two sides). No new substance is created, only distinctions are added to the otherwise formless background. Just as the null set is in-filled with created elements, so the formless substance of the blob is in-formed ("in-formed" as in "information") with distinctions. As more and more distinctions are made (think of drawing more lines on the canvas), we take a snapshot of all the distinct elements (think of regions on the canvas) created by the distinctions so far as the "blocks" in a discrete partition. However, at this point, we must "correct" the concept of a discrete partition as a partition of singleton blocks of elements of U. A discrete partition should be thought simply as the set U itself. Just as we can look backwards from a set U to see all the subsets that might have been intermediary between $\varnothing$ and U, so now we look backward from U

($= \widehat{0}$) to all the possible intermediate steps of distinctions made between the blob $\widehat{1}$ and U. Each step can then be seen as a partition on U, with each block being an element distinguished from the other blocks at that step, but then each block might undergo further distinctions (splittings or fissions) before yielding the blocks = elements of the discrete partition U. The elements of a block in an intermediary partition are only its successors in the fissioning or splitting as more and more distinctions are made.



Two ways to create a set U: "In the beginning was the Void" or "In the beginning was the Blob"

The lattice $\Pi(U)$ of partitions on U consists of all the possible ways that distinctions could have been made starting with the distinctionless blob and ending with the distinctions of the set (= discrete partition) U. Thus the basic and simple notion of a set can be thought of as being 'created' in two dual ways: in-filling a substance-less void or in-forming a formless substance. At any point in either process of creation, we can take a snapshot of obtain a set U and then study all the possible ways it could have been created in the power-set lattice $\mathcal{P}(U)$ or the partition lattice $\Pi(U)$. A 'way-station' between the null set and U would be a subset $\varnothing \hookrightarrow S \hookrightarrow U$ while a way-station between the blob and U would be a partition $U \stackrel{\pi}{\to} X \to 1$.



For a partition (or equivalence relation) $U \stackrel{\pi}{\to} X \to 1$, the customary view is to start with U and then identify elements into equivalence classes or blocks to form the partition. We are suggesting that partitions should be seen the other way around as resulting from making distinctions in the blob. The customary view of seeing partitions arise by starting with U and identifying elements (i.e., throwing away distinctions) is analogous to thinking about a subset S of U by starting with U and then considering the elements that have to be thrown away (the opposite of creation) in order to get the subset S. The dual creation myth puts the emphasis on the distinctions that have to be created or added starting with the blob to get the partition just as we would usually think of a subset in terms of the elements that need to be added to the null set to get the subset. This highlights the duality between distinctions and elements.

The direction of development from blob to U is against the usual direction of the arrows since we might conceive of the categorical setting as the dual category $\mathbf{Set}^{op}$. It may be helpful to rethink this dual category. It is sometimes thought that the morphisms in $\mathbf{Set}^{op}$ are abstract duals to the concrete set functions in **Set**. But the morphisms in $\mathbf{Set}^{op}$ are just as concrete as set functions.

Indeed, they are just set functions as a set of ordered pairs in $X \times Y$ but they are thought of as going in the opposite direction and might be called *cofunctions*.

Set functions $f : X \to Y$ have as their "graph" a binary relation $R_f \subseteq X \times Y$ such that for every $x \in X$, there is an ordered pair $(x, y) \in R_f$ and that ordered pair is unique. If we think of a function $f : X \to Y$ as a process transforming X into Y, then each element $x \in X$ is transformed into some element $f(x) \in Y$ so we might say there are no "deaths" in the transformation. Thus we will say that *elements are preserved* by a function. The duals to elements are distinctions and the opposite to being preserved is being reflected. That is, given any distinction $y \neq y'$ in the codomain, any elements x and x' in X such that (x, y) and (x', y') are in R_f, then $x \neq x'$. This is precisely the other condition on a function, i.e., that it be single-valued. Thus functions are given by the binary relations R_f that preserves elements and reflect distinctions. But there can be fusions (the opposite of a distinction) as when for $x \neq x' \in X$, we have $f(x) = f(x')$. Moreover there can be "births" in the codomain Y, i.e., elements $y \in Y$ that did not come from any element of X via the function f. The two special types of set functions are injections (i.e., monomorphisms in **Set**) where there are no fusions, and surjections (i.e., epimorphisms in **Set**) where there are no births. In terms of elements and distinctions, injections preserve distinctions and surjections reflect elements.

Everything is reversed for the dual notion of a *cofunction*. Every binary relation $R \subseteq X \times Y$ has its opposite $R^* \subseteq Y \times X$ with the same ordered pairs in reversed order. The dual of "no deaths" is "no births". If one thinks of a cofunction as transforming Y into X, then every element of X comes from some (unique) element in Y so there "no births" in X. But there could be "deaths", i.e., elements of Y which do not transform into anything in X (the elements of Y not in the image of f). What looked like fusion (the death of a distinction) in the transformation from X to Y now becomes fission or the creation of a distinction in the transformation in the opposite direction. An element $y \in Y$, might have many descendents resulting from fission (i.e., the elements of $f^{-1}(y)$). The two special types of cofunctions are those that have no deaths (the duals of epis which satisfy the categorical definition of monos in $\mathbf{Set}^{op}$) and those that have no fission (the duals of monos which satisfy the categorical definition of epis in $\mathbf{Set}^{op}$).

Since 1 is "the" terminal object in **Set**, there is a unique epi $X \to 1$ from each set X to 1. We now rethink that function as a cofunction. Distinctions are made in the blob to create the set X (view the epi $X \to 1$ "cofunctionally" as fissions, splittings, or distinctions going from 1 to X) and then further distinctions (with no deaths) are made to create the elements of U so that in "retrospect" the elements of X become "blocks" containing its descendent elements of U. Thus we think of a partition $U \to X \to 1$ being created in a dual manner to a subset $\varnothing \hookrightarrow S \hookrightarrow U$.

1.3 Duality of Elements and Distinctions

The purpose of this section is to argue for "distinctions" as the dual to "elements" in the partition-subset duality. The picturesque language of the births and deaths of elements and distinctions can be formulated in a more sophisticated manner by considering the category of binary relations **rel** between sets. The objects in the category are sets and the morphisms are binary relations $R \subseteq X \times Y$ which are thought of in this case as going "from X to Y". Given two binary relations $R \subseteq X \times Y$ and $S \subseteq Y \times Z$, their composite $R \cdot S \subseteq X \times Z$ is defined by: $(x, z) \in R \cdot S$ if there is a $y \in Y$ such that $(x, y) \in R$ and $(y, z) \in S$.[5] The identity relation on a set X is given by the diagonal $\Delta_X \subseteq X \times X$. A set function $f : X \to Y$ in **Set** has the binary relation $R_f = \{(x, f(x) : x \in X\} \subseteq X \times Y$ in **rel** as its graph. The dual function or cofunction f^* in Set^{op} has the opposite binary relation $R_f^* = \{(y, x) : (x, y) \in R_f\} \subseteq Y \times X$ as its "graph" in **rel** so that we may now consider the compositions $R_f \cdot R_f^*$ and $R_f^* \cdot R_f$.

[5] We are here following the custom of writing the composition of binary relations $X \overset{R}{\to} Y \overset{S}{\to} Z$ as $R \cdot S$ ("R followed by S") whereas the order of composition of general morphisms is usually written in category theory as "$S \cdot R$" ("S follows R").

The set of homomorphisms between two sets in **rel**, $Hom_{\mathbf{rel}}(X, Y)$, also allows a notion of morphism between relations, namely inclusion between binary relations as subsets of $X \times Y$. Thus **rel** is a "bicategory" where we may define an "adjunction" using the analogy that thinks of the sets as categories, the binary relations as functors, and the inclusion relations between binary relations as natural transformations.[6] Then the analogy with the definition of an adjunction using its unit and counit[7], would define an adjunction in **rel** as two binary relations $F \subseteq X \times Y$ and $G \subseteq Y \times X$ such that $\Delta_X \subseteq F \cdot G$ (the "unit") and $G \cdot F \subseteq \Delta_Y$ (the "counit").

The unit condition $\Delta_X \subseteq F \cdot G$ means that for every $x \in X$, there is some $y \in Y$ such that $(x, y) \in F$ and $(y, x) \in G$ so that F as a relation from X to Y is everywhere defined. The counit condition $G \cdot F \subseteq \Delta_Y$ means that if for any $(y, x) \in G$, if there is a $(x, y') \in F$, then $y = y'$. Hence F is single-valued so F is the graph of a function $f : X \to Y$, i.e., $F = R_f$. The unit condition implies that $F^* \subseteq G$ and the counit condition implies $G \subseteq F^*$ so the unique right adjoint to any left adjoint F is F^*. Hence every adjunction in **rel** arises from the graph of a function and its opposite (cofunction). Conversely, for any function $f : X \to Y$, the everywhere-defined condition, $\Delta_X \subseteq R_f \cdot R_f^*$, gives the unit condition, and the single-valued condition, $R_f^* \cdot R_f \subseteq \Delta_Y$, gives the counit condition of an adjunction in **rel**. In other words, adjoints in **rel** are functions.[8]

The duality between monomorphisms and epis and between elements and distinctions is summarized in the following table where the elements of X are represented by the diagonal $\Delta_X \subseteq X \times X$ and the distinctions on X are represented by the complementary set $X \times X - \Delta_X$ and similarly for Y.

$R \subseteq X \times Y$	Action of $R : X \to Y$
Morphism	1. $\Delta_X \subseteq R \cdot R^*$ means:
1.+ 2. =	All X elements preserved by R, and
R is a function	2. $R^* \cdot R \subseteq \Delta_Y$, i.e.,
	$Y \times Y - \Delta_Y \subseteq Y \times Y - R^* \cdot R$ means:
	All Y distinctions reflected by R.
3. = R injective	3. $R \cdot R^* \subseteq \Delta_X$, i.e.,
	$X \times X - \Delta_X \subseteq X \times X - R \cdot R^*$ means:
	All X distinctions preserved by R
4. = R surjective	4. $\Delta_Y \subseteq R^* \cdot R$ means:
	All Y elements reflected by R.

What does it mean to say that all X elements are preserved by R (condition 1)? An element $x \in X$ is taken to Y by the relation R if there is a $y \in Y$ such that $(x, y) \in R$. In that case, $(y, x) \in R^*$ so $(x, x) \in R \cdot R^*$. Hence $\Delta_X \subseteq R \cdot R^*$ means that all the X elements are preserved by R where we may take an element $x \in X$ as being represented by the diagonal element (x, x). To understand the second condition that all Y distinctions are reflected by R, we start with a Y distinction represented by an ordered pair (y, y') where $y \neq y'$. The distinction is reflected back to X by R if whenever that are $(x, y) \in R$ and $(x', y') \in R$, then $x \neq x'$. But if a Y distinction (y, y') was contained in $R^* \cdot R$, then it would mean that the was an $x \in X$ with $(y, x) \in R^*$ and $(x, y') \in R$, i.e., that the Y distinction was not reflected by R. Hence to say that R reflects all Y distinctions means that all the Y distinctions, $Y \times Y - \Delta_Y$, are outside of $R^* \cdot R$, i.e., $Y \times Y - \Delta_Y \subseteq Y \times Y - R^* \cdot R$. These two conditions, where "elements" and "distinctions" have symmetrical dual roles define an adjunction in the category of binary relations **rel** and thus characterize the graphs of functions.

The table is completed by the two other symmetrical conditions that define injections (monos in **Set**) and surjections (epis in **Set**). Given a distinction in X represented by an ordered pair $(x, x') \in X \times X - \Delta_X$, it is transmitted or preserved by R if whenever (x, y) and (x', y') are both in

[6] For more on adjoint functors or adjunctions, see [11] or [5].
[7] See condition (v) in Theorem 2 of [11, p. 81].
[8] For more along these lines, see [19].

R, then $y \neq y'$. If $y = y'$ in this situation, then $(x, x') \in R \cdot R^*$ so all the X distinctions are preserved when they are outside of $R \cdot R^*$, i.e., $X \times X - \Delta_X \subseteq X \times X - R \cdot R^*$ (condition 3 in the table for R being injective). And finally consider any element $y \in Y$ represented by $(y, y) \in \Delta_Y$. It is reflected back to X by R if there is an $x \in X$ with $(x, y) \in R$. In that case $(y, x) \in R^*$so that $(y, y) \in R^* \cdot R$. Hence the condition 4 that all Y elements are reflected by R (R being surjective) means that all Y elements are in $R^* \cdot R$, i.e., $\Delta_Y \subseteq R^* \cdot R$.

Some pains have been taken in this section to argue that the notion of "distinctions" has a role that is dual to the notion of "elements". Ordinary logic is based on the notions of (sub)sets and their elements (generalized in categorical logic to subobjects and their elements). Hence in order to develop a dual "cologic" where partitions are dual to subsets, it has been incumbent on us to develop the notion of a distinction as the dual to the notion of an element. Intuitively we think of an element of a set as an "it." In contrast, a distinction is a "bit" (or rather a "dit") rather than an "it." The distinction between x and x' is the fact that $x \neq x'$.[9] But for mathematical purposes we may represent a distinction by a pair of distinct "things" or elements such as the ordered pair (x, x').[10]

In sum, starting with the notion of a set, we have argued that the notion of "subset of a set" is dual to "partition on a set," and that "elements of a set" is dual to "distinctions between elements of a set."

The conventional logic of subsets leads into probability theory where numerical values are assigned to subsets as events in a sample space. In a similar manner, the logic of partitions is associated with a theory that assigns numerical values to partitions, and that theory is a "logical" version of Claude Shannon's *information theory* [17].

2 Logical Information Theory

2.1 The Space of Ordered Pairs of a Set

The dual creation myth for sets as discrete partitions introduces the main theme in the logic of partitions–*the making of distinctions*. Instead of imagining fully formed elements or "its" being created starting with the null set, the dual creation myth proceeds by adding distinctions to a formless background substance. Once some distinctions have been made, then we look backwards and conceptualize the distinctive elements as being members of a block in a partition. But it would be better to think of the "block" as an element made by some earlier distinctions but before the later distinctions that created the distinctive elements or descendents of the block. The whole development $U \to X \to 1$ can be thought of "cofunctionally" (i.e., going backward against the direction of the arrows in $U \to X \to 1$) as a rooted tree with the root at the blob and the elements of U as the leaves. The "partition" X is a set of nodes in the tree such that each leaf $u \in U$ is the descendent of one and only one node in X.

There is an old philosophical duality between substance and form. We have seen a version of this duality emerge from the category-theoretic duality between monomorphisms and epimorphisms—in terms of sets, subsets of a set and partitions on a set. In the one myth, elements are fully formed and are created by gaining existence or substance. In the dual myth, the substance (the blank canvas) is already there but is formless and then elements are created by making distinctions. In the one

[9] Distinctions can be used to define the number of elements in a finite set, e.g., a set S has three elements if $\exists x, y, z \in S, (x \neq y) \,\&\, (y \neq z) \,\&\, (x \neq z) \,\&\forall w \in S, (w = x) \vee (w = y) \vee (w = z)$.

[10] In economics, there is a basic distinction between rivalrous goods (where more for one means less for another) such a material things ("its") in contrast to non-rivalrous goods (where what one person acquires does not take away from another) such as ideas, knowledge, and information ("bits"). In that spirit, an element of a set represents a material thing, an "it", while the dual notion of a distinction or "dit" represents the immaterial notion of two "its" being distinct, the notion that is the most basic unit of information. The technical relationship between dits and Shannon's bits is explained later. Of course, all mathematical notions are immaterial which is why it may be helpful to look at the real-world contrast between rivalrous and non-rivalrous goods (its and bits).

case, elements are created by in-filling the void with fully formed elements, and, in the other case, elements are created by in-forming an already existing substance. In broad philosophical terms, it is the duality between matter and information.

Claude Shannon's classic 1948 articles [17] developed a statistical theory of communications that is ordinarily called "information theory." Shannon built upon the work of Ralph Hartley [7] twenty years earlier. To make the distinction, the logical analysis of information in terms of distinctions might be called "logical information theory."

The basic conceptual unit in logical information theory is the distinction or *dit* (from "DIsTinction"). Looking backwards from a discrete partition or set U, a partition $\pi : U \to X$ distinguishes certain pairs of elements of U and identifies others. A pair (u, u') of distinct elements of U are distinguished by π, i.e., form a dit of π, if u and u' are in different blocks of π.[11] A pair (u, u') are identified by π and form an *indit* (from INDIsTinction or "identification") of the partition if they are contained in the same block of π. A partition on U can be characterized by either its dits or indits (just as a subset S of U can be characterized by the elements added to the null set to arrive at S or by the elements of U thrown out to arrive at S). When a partition π is thought of as an equivalence relation, then the equivalence relation, as a set of ordered pairs contained in $U \times U = U^2$, is the *indit set* $\mathrm{indit}(\pi)$ of indits of the partition. But from the view point of logical information theory, the focus is on the distinctions, the complementary *dit set* $\mathrm{dit}\,(\pi)$ of dits where $\mathrm{dit}\,(\pi) = (U \times U) - \mathrm{indit}\,(\pi) = \mathrm{indit}\,(\pi)^c$. Rather than think of the partition as resulting from identifications made to the elements of U, we think of it as being formed by making distinctions starting with the blob as in the dual creation myth. The distinctions of π are then measured in terms of the elements created by the later distinctions that formed the universe set U.

There is a natural ("built-in") closure operation on U^2. A subset $C \subseteq U^2$ is *closed* if it contains the diagonal $\{(u, u) \mid u \in U\}$, if $(u, u') \in C$ implies $(u', u) \in C$, and if (u, u') and (u', u'') are in C, then (u, u'') is in C. Thus the closed sets of U^2 are the reflexive, symmetric, and transitive relations, i.e., the equivalence relations (partitions) on U. The intersection of closed sets is closed and the intersection of all closed sets containing a subset $S \subseteq U^2$ is the *closure* $\overline{S}$ of S.

It should be carefully noted that the closure operation on U^2 is not a *topological* closure operation in the sense that the union of two closed set is not necessarily closed. In spite of the closure operation not being topological, we may still refer to the set complements of closed sets as being *open* sets (i.e., the dit sets of partitions). As usual, the interior $\mathrm{int}(S)$ of any subset S is defined as the complement of the closure of its complement: $\mathrm{int}(S) = \left(\overline{S^c}\right)^c$.

The closed sets of $U \times U$ ordered by inclusion form a lattice isomorphic to the lattice $\Pi(U)$ of partitions on U. Given two partitions π and π' on U, the closed set corresponding to the *meet* $\pi \wedge \pi'$ of the partitions is the intersection of their indit sets, i.e.,

$$\mathrm{indit}(\pi \wedge \pi') = \mathrm{indit}\,(\pi) \cap \mathrm{indit}\,(\pi').$$

The closed set corresponding to their *join* $\pi \vee \pi'$ is the closure of the union of their indit sets, i.e.,

$$\mathrm{indit}(\pi \vee \pi') = \overline{\mathrm{indit}\,(\pi) \cup \mathrm{indit}\,(\pi')}.$$

A partition π refines a partition π', $\pi \preceq \pi'$, if $\mathrm{indit}\,(\pi) \subseteq \mathrm{indit}\,(\pi')$. The closed set corresponding to the top or blob $\widehat{1}$ is the whole space of ordered pairs $U \times U$ and the closed set corresponding to the discrete partition or bottom $\widehat{0}$ is the diagonal $\Delta_U = \{(u, u) : u \in U\} \subseteq U \times U$.

Dualizing, the lattice of open sets (dit sets of partitions) of $U \times U$ ordered by inclusion is isomorphic to the dual $\Pi(U)^{op}$. From the view point of logical information theory—which views a partition in terms of its distinctions—it would be more natural to picture the lattice of partitions in this dual way with the discrete partition as the top and the indiscrete partition or blob as the bottom. Then the partition on the "large" end of the inequality sign $\preceq$ would be the more refined

[11] One might also develop the theory using unordered pairs $\{u, u'\}$ but the later development of the theory using probabilistic methods is much facilitated by using ordered pairs (u, u').

partition rather than the less refined partition. Gian-Carlo Rota often quipped that the usual way of writing the "refinement" relation should be called "unrefinement." Some writers [6] have defied the tradition, defined the refinement relation the other way, and made the other necessary changes such as taking what is usually defined as the meet of two partitions as their join. But after registering this protest, we will stick to the conventional treatment of the lattice of partitions so the lattice of open dit sets is isomorphic to the dual $\Pi(U)^{op}$.

We have taken some pains to emphasize two different conceptual origins of the notion of a set in the two dual creation myths. This does not preclude intertranslation between subset concepts and partition concepts. The isomorphism between the lattice of closed subsets of $U \times U$ and the lattice of partitions shows how all the information in the partition lattice can be translated back into the subsets and closure relation. Even the notion of a characteristic function $\chi_S : U \to 2 = \{0, 1\}$ of a subset $S = \chi_S^{-1}(1) \subseteq U$ carries over. Consider $2^2 = \{(0,0), (1,1), (0,1), (1,0)\}$ where the subset $\{(0,0), (1,1)\}$consisting of the self-pairs is a closed set, the only closed set different from the whole set. Then given any closed set $C \subseteq U^2$, define the map $\chi_C((u,u')) = (0,x)$ where $x = 0$ if $(u,u') \in C$ and $x = 1$ if $(u,u') \notin C$. Then $\chi_C^{-1}((0,0)) = C$ and, indeed, $\chi_C : U^2 \to 2^2$ is a *closed map* in the sense that the inverse image of closed sets are closed. Thus the partitions on U are in 1-1 correspondence with the closed maps $U^2 \to 2^2$ where the first component always maps to 0.

2.2 Some Set Structure Theorems

After Shannon's information theory was presented, there was a spate of new definitions of "entropy" with various properties but without concrete interpretations. Here we are taking an opposite approach of starting with an interpretation that arises naturally in the development of a dual logic for partitions. The basic notion of a distinction or dit is then seen as the logical unit of information and a "logical information theory" can be developed based on that concrete interpretation. When the universe set U is finite, then we have a numerical notion of "information" or "entropy" of a partition π in the number of distinctions $|\mathrm{dit}(\pi)|$ particularly when normalized by the number of ordered pairs, i.e., $|\mathrm{dit}(\pi)| / |U|^2$. This logical "number of dits" notion of information or entropy can then be related to Shannon's measure of information or entropy. Before restricting ourselves to finite U, there are a few structure theorems that are independent of cardinality.

The unit of information is the dit so the information in a partition π is its dit set $\mathrm{dit}(\pi)$. The information common to two partitions π and σ, their *mutual information set*, would be the intersection of their dit sets (which is not necessarily the dit set of a partition):

$$\mathrm{Mut}(\pi, \sigma) = \mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma).$$

The dit set of the blob $\widehat{1}$ is the empty set since it distinguishes nothing (and thus has no information) while the dit set of the discrete partition $\widehat{0}$ is all ordered pairs except the diagonal $\Delta_U = \{(u,u) : u \in U\}$. Shannon deliberately defined the notion of information so that it would be "additive" in the sense that the measure of information in two independent probability distributions would be the sum of the information measures of the two separate distributions.[12] But this is not true at the logical level with information defined as distinctions. There is *always* mutual information between two non-blob partitions.

Proposition 1 *Given two partitions π and σ on U with $\pi \neq \widehat{1} \neq \sigma$, $\mathrm{Mut}(\pi, \sigma) \neq \varnothing$.*

Since π is not the blob, consider two elements u and u' distinguished by π but identified by σ [otherwise $(u,u') \in \mathrm{Mut}(\pi,\sigma)$]. Since σ is also not the blob, there must be a third element u'' not in the same block of σ as u and u'. But since u and u' are in different blocks of π, the third

[12] Two partitions π and σ are (stochastically) *independent* if the probability distributions $\{p_B\}_{B \in \pi}$ and $\{p_C\}_{C \in \sigma}$ are independent in the sense that: $\frac{|B \cap C|}{|U|} = p_{B \cap C} = p_B p_C = \frac{|B||C|}{|U \times U|}$ for all $B \in \pi$ and $C \in \sigma$.

element u'' must be distinguished from one or the other or both in π. Hence (u, u'') or (u', u'') must be distinguished by both partitions and thus be in their mutual information set $\mathrm{Mut}(\pi, \sigma)$.

Just as the union of two closed sets is not necessarily closed, the intersection of two open sets is not necessarily open. However the interior of the mutual information set is the dit set of the join partition:

$$\mathrm{int}\left(\mathrm{Mut}(\pi,\sigma)\right) = \overline{\left[\left(\mathrm{dit}\left(\pi\right) \cap \mathrm{dit}\left(\sigma\right)\right)^c\right]}^c = \overline{\left[\mathrm{indit}\left(\pi\right) \cup \mathrm{indit}\left(\sigma\right)\right]}^c = \mathrm{indit}\left(\pi \vee \sigma\right)^c = \mathrm{dit}\left(\pi \vee \sigma\right).$$

While the mutual information set of two non-trivial partitions cannot be empty, the interior can be empty, and that is equivalent to $\pi \vee \sigma = \widehat{1}$.

It is easy to characterize the structure of the relevant subsets of U^2 using the usual notions of blocks of a partition. Given a partition π on U with blocks $\{B\}_{B\in\pi}$, let $B \times B'$ be the Cartesian product of B and B'. Then

$$\begin{aligned} \mathrm{indit}\left(\pi\right) &= \bigcup_{B\in\pi} B \times B, \text{ and} \\ \mathrm{dit}\left(\pi\right) &= \bigcup_{\substack{B\neq B' \\ B,B'\in\pi}} B \times B' = U \times U - \mathrm{indit}\left(\pi\right). \end{aligned}$$

The mutual information set can also be characterized in this manner.

Proposition 2 *Given partitions π and σ with blocks $\{B\}_{B\in\pi}$ and $\{C\}_{C\in\sigma}$, then*

$$\mathrm{Mut}\left(\pi,\sigma\right) = \bigcup_{B\in\pi, C\in\sigma} \left(B - \left(B \cap C\right)\right) \times \left(C - \left(B \cap C\right)\right).$$

The union (which is a disjoint union) will include the pairs (u, u') where for some $B \in \pi$ and $C \in \sigma$, $u \in B - (B \cap C)$ and $u' \in C - (B \cap C)$. Since u' is in C but not in the intersection $B \cap C$, it must be in a different block of π than B so $(u, u') \in \mathrm{dit}\left(\pi\right)$. Symmetrically, $(u, u') \in \mathrm{dit}\left(\sigma\right)$ so $(u, u') \in \mathrm{Mut}\left(\pi, \sigma\right) = \mathrm{dit}\left(\pi\right) \cap \mathrm{dit}\left(\sigma\right)$. Conversely if $(u, u') \in \mathrm{Mut}\left(\pi, \sigma\right)$ then take the B containing u and the C containing u'. Since (u, u') is distinguished by both partitions, $u \notin C$ and $u' \notin B$ so that $(u, u') \in (B - (B \cap C)) \times (C - (B \cap C))$.

2.3 Logical Information Theory on Finite Sets

For a finite set U, the numerical "dit count" measure of information can be defined and compared to Shannon's measure for finite probability distributions. Since the information set of a partition π on U is its set of distinctions $\mathrm{dit}\left(\pi\right)$, the un-normalized numerical measure of the information of a partition is simply the cardinality of that set, $|\mathrm{dit}\left(\pi\right)|$. But the description of the information in π is all relative to freezing the development of the partitions with the discrete partition U so our main focus is on the normalized *logical information content* or *logical entropy* of a partition π which is:

$$\boxed{h\left(\pi\right) = \frac{|\mathrm{dit}(\pi)|}{|U|^2}}.$$

Probability theory started with the finite case where there was a finite set U of possibilities (the finite sample space) and an event was a subset $S \subseteq U$. Under the Laplacian assumption that each outcome was equiprobable, the probability of the event S was:

$$\mathrm{Prob}\left(S\right) = \frac{|S|}{|U|}.$$

This is the probability that any randomly chosen element of U would exist in the subset S. In view of the dual relationship between existing in a subset and being distinguished by a partition (i.e., the duality between existence and information), the analogous concept would be the probability that an ordered pair (u, u') of elements of U chosen randomly (with replacement) would be distinguished by a partition π. That is precisely the *probabilistic interpretation of the logical information* of a partition $h(\pi) = |\mathrm{dit}(\pi)| / |U|^2$ (since each pair randomly chosen from $U \times U$ is equiprobable).

In view of the close analogy between subsets of a set and partitions on a set where the partitions are treated in terms of their dit sets, we have a bridge to carry over techniques in finite probability theory to partitions. Since $\mathrm{dit}(\pi \wedge \sigma) = \mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma)$, the probability that an ordered pair randomly chosen (always with replacement) is distinguished by either a partition π or a partition σ (or both) is the probability $|\mathrm{dit}(\pi \wedge \sigma)| / |U|^2$. Since that is the information content or entropy $h(\pi \wedge \sigma) = |\mathrm{dit}(\pi \wedge \sigma)| / |U|^2$ for the meet $\pi \wedge \sigma$, the entropy of the meet has the simple probabilistic interpretation as the probability that a random pair would be distinguished by one or the other of the partitions. The probability that a randomly chosen pair would be distinguished by both partitions π and σ would be given by the relative cardinality of the mutual information set which is called the *mutual information* of the partitions:

$$m(\pi, \sigma) = \frac{|\mathrm{Mut}(\pi,\sigma)|}{|U|^2}.$$

Since the cardinality of intersections of sets can be analyzed using the inclusion-exclusion principle, we have:

$$|\mathrm{Mut}(\pi, \sigma)| = |\mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma)| = |\mathrm{dit}(\pi)| + |\mathrm{dit}(\sigma)| - |\mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma)|.$$

Normalizing, the probability that a random pair is distinguished by both partitions is given by the "modular law":

$$m(\pi, \sigma) = \frac{|\mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma)|}{|U|^2} = \frac{|\mathrm{dit}(\pi)|}{|U|^2} + \frac{|\mathrm{dit}(\sigma)|}{|U|^2} - \frac{|\mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma)|}{|U|^2} = h(\pi) + h(\sigma) - h(\pi \wedge \sigma).$$

This can be extended by the inclusion-exclusion principle to any number of partitions. Since the dits of the join are obtained as the interior of the mutual information set $\mathrm{Mut}(\pi, \sigma)$, the information contents of the join and meet satisfy the:

$$\text{Submodular inequality: } h(\pi \vee \sigma) + h(\pi \wedge \sigma) \leq h(\pi) + h(\sigma).$$

2.4 Using Probabilistic Methods

Since the logical entropy of a partition on a finite set can be given a direct probabilistic interpretation, it is not surprising that many methods of probability theory can be directly harnessed to develop the theory. Each partition π on a finite set U defines a probability distribution:

$$p_B = \frac{|B|}{|U|} \text{ for blocks } B \in \pi.$$

Since there are no empty blocks, $p_B > 0$ and $\sum_{B \in \pi} p_B = 1$. Since the dit set of a partition is $\mathrm{dit}(\pi) = \bigcup_{B \neq B'} B \times B'$, its size is $|\mathrm{dit}(\pi)| = \sum_{B \neq B'} |B|\,|B'|$. Thus the logical information or entropy in a partition as the normalized size of the dit set can be developed as follows:

$$h(\pi) = \frac{\sum_{B \neq B'} |B|\left|B'\right|}{|U| \times |U|} = \sum_{B \neq B'} p_B p_{B'} = \sum_{B \in \pi} p_B (1 - p_B) = 1 - \sum_{B \in \pi} p_B^2.$$[13]

[13] At about the same time as Shannon's 1948 paper on information theory, Edward H. Simpson, a British statistician, (independently) proposed $\sum_{B \in \pi} p_B^2$ as a measure of species concentration (the opposite of diversity) where π is the partition of animals or plants according to species and where each animal or plant is considered as equiprobable. And Simpson gave the important interpretation of this measure as "the probability that two individuals chosen at random and independently from the population will be found to belong to the same group."[18, p. 688] Hence $1 - \sum_{B \in \pi} p_B^2$ is the probability that a random ordered pair will belong to different groups. This early appearance of the formula for logical entropy is known as "Simpson's index of diversity."

We define the *partition indicator function* $I_\pi : U \to [0,1]$ of a partition π as:

$$I_\pi(u) = p_B \text{ where } u \in B.$$

Using the linear expectation operator $E[-]$ on random variables defined on U, we have:

$$h(\pi) = \sum_{B \in \pi} p_B (1 - p_B) = E[1 - I_\pi] = 1 - E[I_\pi].$$

Thus the logical entropy $h(\pi)$ not only has a direct probabilistic interpretation (the probability that a randomly chosen pair is distinguished by the partition), it can also be interpreted as the average of the values of the complementary probabilities $1 - p_B$ on the blocks. This "average-of-complementary-probabilities" version of the logical entropy is important to see the connection to Shannon's measure of information content. Having defined and interpreted logical entropy in terms of the distinctions of a set partition, we may if desired "kick away the ladder" and define it directly for any finite probability distribution $p = \{p_1, ..., p_n\}$ as $h(p) = \sum_{i=1}^n p_i (1 - p_i) = 1 - \sum_{i=1}^n p_i^2$.

3 The Connection Between the Logical and Shannon's Information Theories

3.1 Comparing the Logical and Shannon Entropies

This result allows us to visualize the connection between the logical entropy $h(\pi) = \sum_{B \in \pi} p_B (1 - p_B)$ of a partition and Shannon's entropy $H(\pi) = \sum_{B \in \pi} p_B \log_2(\frac{1}{p_B})$ of the partition. But first we must motivate Shannon's entropy concept. We have taken some pains to carefully distinguish between subset-based reasoning in terms of elements of subsets of U and partition-based reasoning in terms of the ordered pairs of distinctions made by a partition on U. For instance, we had the contrast between the probability $\text{Prob}(S) = \frac{|S|}{|U|}$ of a random element being in a subset S and the probability $h(\pi) = \frac{|\text{dit}(\pi)|}{|U|^2}$ of a random pair being distinguished by a partition π. Shannon, in effect, uses subset-based reasoning (shared with Hartley) to arrive at a notion of entropy for a block of a partition—where the block is considered as a subset of U—and then the block values are averaged over all the blocks in a partition to get a partition value. Hartley and Shannon start with the question of the information required to single an element u out of a set U, e.g., to single out the sent message out of a set of possible messages. Alfred Renyi has also emphasized this "search-theoretic" approach to information theory (see [13] or numerous papers in [15]).

One intuitive measure of that information would just be the cardinality $|U|$ of the set, and, as we will see, that is indeed a multiplicative version of Shannon's entropy. But Hartley and Shannon wanted the additivity that comes from taking the logarithm of the set size $|U|$. If $|U| = 2^n$ then this allows the crucial Shannon interpretation of $\log_2(|U|) = n$ as the minimum number of yes-or-no questions it takes to single out any designated element of the set. In a mathematical version of the game of twenty questions (like Rényi's Hungarian game of "Bar-Kochba"), think of each element of U as being assigned a unique binary number with n digits. Then the minimum n questions can just be the questions asking for the i^{th} binary digit of the hidden designated element. Each answer gives one *bit* (short for "binary digit") of information. With this motivation for the case of $|U| = 2^n$, Shannon takes $\log(|U|)$ (logs are always to the base 2 unless otherwise indicated) as the measure of the information required to single out a hidden element in set with $|U|$ elements.

The next step is to break down the finding of the hidden element into two steps; finding a subset B containing the element and then finding the element within the subset. Assuming that the information to single out B and to single out the element within B is additive, then the additive measure of information $H(B)$ to single out B would satisfy the equation: $\log(|U|) = H(B) + \log(|B|)$. Then we can solve for the information to single out a subset B as: $H(B) = \log(|U|) -$

$\log(|B|) = \log\left(\frac{|U|}{|B|}\right) = \log\left(\frac{1}{p_B}\right)$. The basic idea is that singling out or distinguishing one element from a set of equally likely elements is the log of the number of elements. Hence to measure the information obtained in singling out a subset B out of U, it is thought of as being singled out of a set of equal subsets, where there are (as it were) $\frac{|U|}{|B|}$ subsets of size $|B|$ in U.[14] Hence the information obtained in distinguishing a subset B out of $\frac{|U|}{|B|}$ equal subsets is $\log\left(\frac{|U|}{|B|}\right)$. Then a partition π is brought in as a set of mutually exclusive and jointly exhaustive subsets $\{B\}_{B\in\pi}$. Since the hidden element must be contained in one and only one block, the information or entropy of the partition could be defined as the average of the block values:

$$\text{Shannon's entropy: } H(\pi) = \sum_{B\in\pi} p_B H(B) = \sum_{B\in\pi} p_B \log\left(\frac{1}{p_B}\right).$$

It should be carefully noted that the interpretation of the Shannon entropy is for the block, and the measure is extended to a partition only by averaging over the blocks.

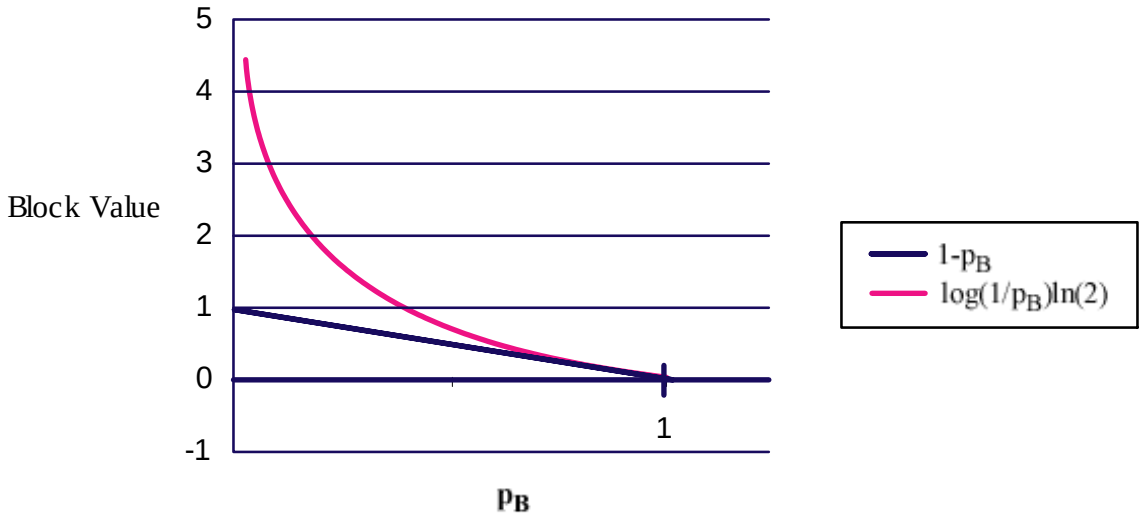
This can be directly compared to the logical entropy $h(\pi) = \sum_{B\in\pi} p_B(1-p_B)$ which arose from quite different partition-based reasoning (e.g., with the search of a single designated element playing no role). Nevertheless, the formula $\sum_{B\in\pi} p_B(1-p_B)$ can be viewed as an average over the quantities which play the role of "block values" $h(B) = (1-p_B)$. This "block value" cannot be directly interpreted as a (normalized) dit count since there is no such thing as the dit count for a block. The dits are the pairs of elements in distinct blocks. However, we could mimic Shannon's reasoning about a block B by considering a similarly hypothetical partition with $\frac{|U|}{|B|}$ equal blocks of size $|B|$ and then compute the normalized dit count for that hypothetical partition. Since each block of that hypothetical partition has the same probability $p_B = \frac{|B|}{|U|}$, its logical entropy is the sum of the $\frac{|U|}{|B|} = \frac{1}{p_B}$ quantities $p_B(1-p_B)$ which yields the previous "block value" $\frac{1}{p_B}p_B(1-p_B) = 1-p_B = h(B)$. Then the average of these logical entropies for the hypothetical partitions with $\frac{|U|}{|B|}$ equal blocks would give the logical entropy $h(\pi) = \sum_{B\in\pi} p_B h(B) = \sum_{B\in\pi} p_B(1-p_B)$ of the partition π.

The surprising thing is that the interpretation of the logical entropy (normalized dit counts) survives the averaging even though all the blocks of π might have different sizes. The interpretation "commutes" with the averaging of block values. In contrast, the Shannon measure of information as the minimum number of binary questions it takes to distinguish the elements of a set of equiprobable elements (or subsets) does not commute with the averaging over the set of different-sized blocks in a partition. The logical entropy of a partition on a set can be interpreted as the actual number of dits (normalized) distinguished by the partition while the Shannon entropy of a partition is always the *expected* number of bits it takes to distinguish the blocks.

One of the important tools used in Shannon's information theory is the logarithmic inequality: for any $x > 0$, $\ln(x) \leq x - 1$ with equality only at $x = 1$ (where ln() is the natural logarithm). This is easily proved by showing that $f(x) = x - 1 - \ln(x)$ has its unique minimum value of $f(1) = 0$. Taking the negative of both sides of the inequality allows us to directly see the relationship at the block level between logical and Shannon's entropy: for $p_B > 0$,

$$\text{logarithmic inequality: } h(B) = 1 - p_B \leq \ln\left(\frac{1}{p_B}\right) = \log_2\left(\frac{1}{p_B}\right)\ln(2) = H(B)\ln(2).$$

[14] We should always say "as it were" here since $|U|/|B|$ need not be an integer unless we rig it that way. But the heuristic argument goes through anyway.



Comparison of Block Values for Logical and Shannon Entropies

Averaging over the block values, we have:

$$0 \leq h(\pi) = \sum_{B \in \pi} p_B h(B) = \sum_{B \in \pi} p_B (1 - p_B) \leq \sum_{B \in \pi} p_B \ln(\tfrac{1}{p_B}) = \ln(2) \sum_{B \in \pi} p_B \log_2\left(\tfrac{1}{p_B}\right) = \ln(2) H(\pi) \leq H(\pi)$$

with equality only when $p_B = 1$ since both entropies take the blob as having zero information.

Since both block values of the entropies are a function of block probability p_B, we can eliminate that variable to obtain a direct relationship between the block values.

$$\boxed{h(B) = 1 - \tfrac{1}{2^{H(B)}} \text{ and } H(B) = \log\left(\tfrac{1}{1-h(B)}\right)}$$

Relationship between logical and Shannon block entropies

$H(B)$ is the minimal number of binary partitions necessary to distinguish $m = 2^{H(B)}$ elements (as it were) such as the blocks in the hypothetical partition with $m = \frac{1}{p_B} = 2^{\log(1/p_B)} = 2^{H(B)}$ equal blocks. The logical entropy $h(B)$ of that partition has the same interpretation as any logical entropy—the probability that a random ordered pair of elements will be distinguished by the partition. Since the m blocks have the same size, the probability that the first draw is from any specific block is $\frac{1}{m}$. Hence the probability that a random ordered pair is distinguished is simply the probability $1 - \frac{1}{m} = 1 - \frac{1}{2^{H(B)}}$ that the second draw is in a different block than the first. Hence the formula $h(B) = 1 - \frac{1}{2^{H(B)}}$ gives a direct and clear intuitive relationship between the two notions of entropy at the level of blocks (or at the level of individual points in a finite sample space when both entropies are defined in terms of a probability distribution on a finite sample space).

To summarize the comparison up to this point, the logical theory and Shannon's theory start by posing different questions which then turn out to be precisely related. Shannon's statistical theory of communications is concerned with determining the sent message out of a set of possible messages. In the basic case, the messages are equiprobable so it is abstractly the problem of determining the hidden designated element out of a set of equiprobable elements which, for simplicity, we can assume has 2^n elements. The process of determining the hidden element (e.g., the sent message) can be conceptualized as the process of asking binary questions which split the set of possibilities into equiprobable parts. The answer to the first question determines which subset of 2^{n-1} elements contains the hidden element and that provides 1 bit of information. An independent equal-blocked

binary partition would split each of the 2^{n-1} element blocks into equal blocks with 2^{n-2} elements each. Thus 2 bits of information would determine which of those 2^2 blocks contained the hidden element, and so forth. Thus n independent equal-blocked binary partitions would determine which of the resulting 2^n blocks contains the hidden element. Since there are 2^n elements each of those blocks is a singleton so the hidden element has been determined. Hence the problem of finding a designated element among 2^n equiprobable elements requires $\log(2^n) = n$ bits of information.

The logical theory starts with the basic notion of a distinction between elements and defines the logical information in a set of distinct 2^n elements as the number of distinctions that need to be made to distinguish the 2^n elements. The distinctions are counted as ordered rather than unordered pairs (in order to better apply the machinery of probability theory) and the number of distinctions or dits is normalized by the number of all ordered pairs. Hence a set of 2^n distinct elements would involve $2^n \times 2^n - 2^n = 2^{2n} - 2^n = 2^n(2^n - 1)$ distinctions which normalizes to $\frac{2^{2n}-2^n}{2^{2n}} = 1 - \frac{1}{2^n}$ and which can be interpreted as the probability that a pair of elements chosen at random and with replacement between the draws will be a pair of *distinct* elements (i.e., the probability that the second element drawn is distinct from the first).

The connection between the two approaches can be seen by computing the total number of distinctions made by intersecting the n independent equal-blocked binary partitions in Shannon's approach. There are only $2^{2n} - 2^n = 2^n(2^n - 1)$ possible distinctions between elements in a set of 2^n elements and if any distinction, say (u, u') for $u \neq u'$, were left unmade, then the Shannon procedure would not be able to determine the hidden designated element if it were either u or u'. Hence the intersection of the n independent equal-blocked binary partitions needs to make all the $2^n(2^n - 1)$ possible distinctions. The first partition which creates two sets of 2^{n-1} elements each thereby creates $2^{n-1} \times 2^{n-1} = 2^{2n-2}$ distinctions as unordered pairs and $2 \times 2^{2n-2} = 2^{2n-1}$ distinctions as ordered pairs. The next binary partition splits each of those blocks into equal blocks of 2^{n-2} elements. Each split block creates $2^{n-2} \times 2^{n-2} = 2^{2n-4}$ new distinctions as unordered pairs and there were two such splits so there are $2 \times 2^{2n-4} = 2^{2n-3}$ additional unordered pairs of distinct elements created or 2^{2n-2} new ordered pair distinctions. In a similar manner, the third partition creates 2^{2n-3} new dits and so forth down to the n^{th} partition which adds 2^{2n-n} new dits. Thus in total, the intersection of the n independent equal-blocked binary partitions has created

$$2^{2n-1} + 2^{2n-2} + ... + 2^{2n-n} = 2^n\left(2^{n-1} + 2^{n-2} + ... + 2^0\right) = 2^n\left(\tfrac{2^n-1}{2-1}\right) = 2^n(2^n - 1)$$

(ordered pair) distinctions which is all the dits on a set with 2^n elements. This is the instance of the block value relationship $h(B) = 1 - \frac{1}{2^{H(B)}}$ when the block B is a singleton in a 2^n element set so that $H(B) = \log\left(\frac{1}{1/2^n}\right) = \log(2^n) = n$ and $h(B) = 1 - \frac{1}{2^n}$. Thus the Shannon entropy as the number of independent equal-blocked binary partitions it takes to single out a hidden designated element in a 2^n element set is *also* the number of independent equal-blocked binary partitions it takes to distinguish all the elements of a 2^n element set from each other.

The connection between Shannon entropy and logical entropy boils down to two steps.

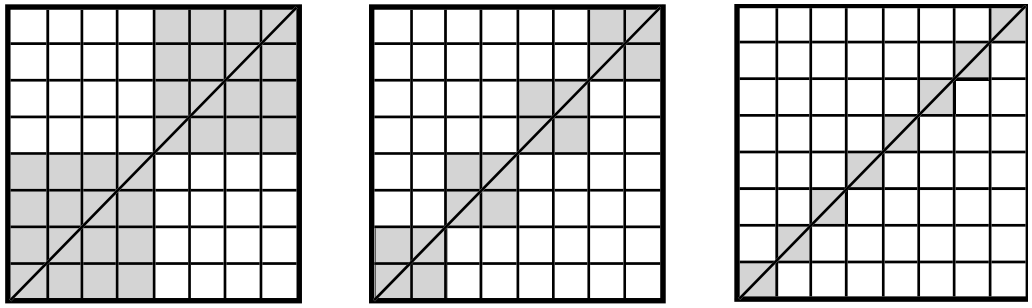
1. The first step uses the basic fact that singling out a hidden element ("sent message") in a set is the same as being able to distinguish any pair of distinct elements (since if a pair was left undistinguished, the hidden element could not be singled out if it were one of the elements in that undistinguished pair). This gives what might be called the *second interpretation* of Shannon entropy as a measure of the information necessary to distinguish between all the distinct messages in the set of possible messages in addition to the usual interpretation as measuring the information necessary to determine the hidden designated element, i.e., the sent message.

2. The second step is that in addition to the Shannon measure of the information necessary to make all the distinction, we may use the logical measure that simply counts the distinctions which is normalized by the total number of ordered pairs of elements.

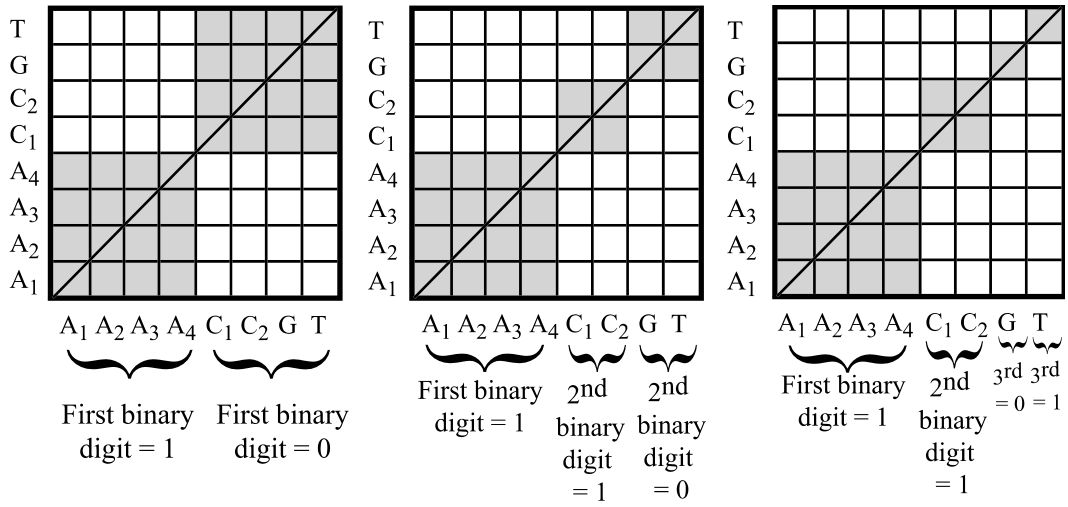
In short,

> *the Shannon measure counts the minimum number of binary partitions to make the distinctions between the elements of a set while the logical measure is the (normalized) count of distinctions themselves.*

Example 1: This connection might be illustrated by considering the case of $n = 3$ so that it takes only 3 independent binary partitions to distinguish the $2^3 = 8$ elements of the set (represented by the 8 diagonal squares) and to efficiently create the $2^3\left(2^3 - 1\right) = 8 \times 7 = 56$ ordered pair distinctions represented by the off-diagonal squares.



Example 2: It might be useful to also consider an example where the blocks are unequal. Consider the four symbols C, A, G, and T in the genetic code and suppose the probabilities are $p_A = 1/2$, $p_C = 1/4$, and $p_G = p_T = 1/8$.[16] We might model this with a set U of eight letters, 4 copies of A, 2 copies of C, and one copy of G and T. There is a (Fano) code for these possibilities, $A = 1$, $C = 01$, $G = 000$, $T = 001$. Then one binary digit will distinguish the block of A's from the other blocks (square on the left below), two binary digits will distinguish the C-block from G and T (the middle square below) and three binary digits will distinguish G from T. But this does not represent 3 bits of information since the blocks were of different probabilities. The block values are: $H(A) = \log\left(\frac{1}{1/2}\right) = 1$, $H(C) = \log\left(\frac{1}{1/4}\right) = 2$, and $H(G) = H(T) = \log\left(\frac{1}{1/8}\right) = 3$, which are the number of symbols in the given code. But the average number of binary digits in the code needed to distinguish the four blocks is: $\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 \times 2 = 1\frac{3}{4}$ bits per block which is the Shannon entropy of the probability distribution.



Now consider the action of the three binary digits in creating dits. The first binary partition (i.e., distinguishing according to the first binary digit) created 32 white cells representing distinctions (square on the left above) so the normalized dit count is $h(A) = \frac{32}{64} = \frac{1}{2} = 1 - \frac{1}{2^{H(A)}}$. The second binary digit distinguished the C block of 2 elements among 8 and a (hypothetical) partition with four equal blocks of 2 elements would have a normalized dit count of $h(C) = \frac{64-(4\times 2\times 2)}{64} = \frac{3}{4} = 1 - \frac{1}{4} = 1 - \frac{1}{2^{H(C)}}$. The third binary digit distinguished the last two singleton blocks and a partition with eight singleton blocks has the logical entropy: $h(G) = h(T) = \frac{64-8}{64} = \frac{7}{8} = 1 - \frac{1}{8} = 1 - \frac{1}{2^{H(G)}} = 1 - \frac{1}{2^{H(T)}}$. When we average these logical entropy block values over the four blocks, the result is:

$$\left(\tfrac{1}{2} \times \tfrac{1}{2}\right) + \left(\tfrac{1}{4} \times \tfrac{3}{4}\right) + 2\left(\tfrac{1}{8} \times \tfrac{7}{8}\right) = \tfrac{8}{32} + \tfrac{6}{32} + \tfrac{7}{32} = \tfrac{21}{32}.$$

It was previously noted that the interpretation commutes with the averaging over block values for logical entropy. Hence that average is the same as the direct count of the 42 white cells in the chessboard on the right above representing the distinctions of the partition so the normalized dit count is the logical entropy of the partition: $\frac{42}{64} = \frac{21}{32}$.

Example 3: The logic of the connection between intersecting independent equal-blocked partitions and efficiently creating dits is not dependent on the choice of base 2. Consider the coin-weighing problem where one has a balance scale and a set of 3^n coins all of which look alike but one is counterfeit (the hidden designated element) and is lighter than the others. The coins might be numbered using the n-digit numbers in mod 3 arithmetic where the three digits are 0, 1, and 2. The n independent ternary partitions are arrived at by dividing the coins into three piles according to the i^{th} digit as $i = 1, ..., n$. To use the n partitions to find the false coin, two of the piles are put on the balance scale. If one side is lighter, then the counterfeit coin is in that block. If the two sides balance, then the light coin is in the third block of coins not on the scale. Thus n weighings (i.e., the intersection of n independent equal-blocked ternary partitions) will determine the n ternary digits of the false coin, and thus the ternary Shannon entropy is $\log_3\left(\frac{1}{1/3^n}\right) = \log_3(3^n) = n$ trits. As before we can interpret the intersecting of independent partitions not only as the most efficient way to find the hidden element (e.g., the false coin or the sent message) but as the most efficient way to make all the distinctions between the elements of the set. The first partition (separating by the first ternary digit) creates 3 equal blocks of 3^{n-1} elements each so that creates $3 \times 3^{n-1} \times 3^{n-1} = 3^{2n-1}$ unordered pairs of distinct elements or $2 \times 3^{2n-1}$ ordered pair distinctions. The partition according to the second ternary digit divides each of these three blocks into three equal blocks of 3^{n-2} elements each so the additional unordered pairs created are $3 \times 3 \times 3^{n-2} \times 3^{n-2} = 3^{2n-2}$ or $2 \times 3^{2n-2}$ ordered pair distinctions. Continuing in this fashion, the n^{th} ternary partition adds $2 \times 3^{2n-n}$ dits. Hence the total number of dits created by intersecting the n independent partitions is:

$$2 \times \left[3^{2n-1} + 3^{2n-2} ... + 3^n\right] = 2 \times \left[3^n\left(3^{n-1} + 3^{n-2} ... + 1\right)\right] = 2 \times \left[3^n \tfrac{(3^n-1)}{3-1}\right] = 3^n(3^n - 1)$$

which is the total number of ordered pair distinctions between the elements of the 3^n element set. Thus the Shannon measure in trits is the minimum number of ternary partitions needed to create all the distinctions between the elements of a set. The base-3 Shannon entropy is $H_3(\pi) = \sum_{B\in\pi} p_B \log_3\left(\frac{1}{p_B}\right)$ which for this example of the discrete partition on a 3^n element set U is $H_3\left(\widehat{0}\right) = \sum_{u\in U} \frac{1}{3^n} \log_3\left(\frac{1}{1/3^n}\right) = \log_3(3^n) = n$ which can also be thought of as the block value entropy for a singleton block so that we may apply the block value relationship. The logical entropy of the discrete partition on this set is: $h\left(\widehat{0}\right) = \frac{3^n(3^n-1)}{3^{2n}} = 1 - \frac{1}{3^n}$ which could also be thought of as the block value of the logical entropy for a singleton block. Thus the entropies for the discrete partition stand in the block value relationship which for base 3 is:

$$h(B) = 1 - \tfrac{1}{3^{H_3(B)}}.$$

The examples help to show how the logical notion of a distinction underlies the Shannon measure of information, and how a complete procedure for finding the hidden element (e.g., the sent message) requires being able to make all the distinctions in a set of elements. But this should not be interpreted as showing that the Shannon's information theory "reduces" to the logical theory. The Shannon theory is addressing an additional question of finding the unknown element. One can have all the distinctions between elements, e.g., the assignment of distinct base-3 numbers to the 3^n coins, without knowing which element is the designated one. Information theory becomes a theory of the *transmission* of information, i.e., a theory of communication, when that second question of "receiving the message" as to which element is the designated one is the focus of analysis. In the coin example, we might say that the information about the light coin was always there in the nature of the situation (i.e., on the sender side) but was unknown to an observer (i.e., on the receiver side). The coin weighing scheme was a way for the observer to elicit the information out of the situation. Similarly, the game of twenty questions is about finding a way to uncover the hidden answer—which was all along distinct from the other possible answers (on the sender side). It is this question of the transmission of information (and the noise that might interfere with the process) that carries Shannon's statistical theory of communications well beyond the bare-bones logical analysis of information in terms of distinctions.

3.2 Multiplicative Shannon Entropy

The fact that the Shannon motivation works for other bases than 2 suggests that there might be a base-free version of the Shannon measure. Sometimes the reciprocal $\frac{1}{p_B}$ of the probability of an event B is interpreted as the "surprise-value information" conveyed by the occurrence of B. But there is a better concept to use than the vague notion of "surprise-value information." For any probability p, we define this reciprocal $\frac{1}{p}$ as the *associated number of (equiprobable) elements* (always "as it were" since it need not be an integer) since that is the number of equiprobable elements in a set so that the probability of choosing any particular element is p. If an outcome in a sample space has probability $p = \frac{1}{n}$, then picking that outcome has the same probability as picking any particular element from a set of n equiprobable elements. For instance, for a block probability $p_B = \frac{|B|}{|U|}$, its associated number of elements or, rather, blocks (in this case) $\frac{|U|}{|B|}$ was the number of blocks in the hypothetical equal-blocked partition with each block like B. It might be said that the "surprise-value information" conveyed when it is learned which element in U is the hidden designated element was the reciprocal $\frac{1}{1/|U|} = |U|$ of its probability, but that is also the associated number of elements to the probability $1/|U|$ since all the elements were equiprobable. In that case the equal-blocked partition where each block is like the singleton block of the designated element is just the discrete partition on U. Our task is to develop this "associated number of blocks" (or "surprise value") measure of information for partitions.

If events B and C were independent, then $p_{B\cap C} = p_B p_C$ so the number of elements associated with the occurrence of both events is the product of the number of elements associated with the separate events. Let $H_m(B)$ be the "number of blocks" information contained in learning that the hidden element was in a subset B. Then the number-of-blocks information contained in learning which element was designated in U could be taken as the product of the information that the element is in the subset B times the information about which element of B was designated, i.e., $|U| = H_m(B)\,|B|$. Thus $H_m(B) = \frac{|U|}{|B|} = \frac{1}{p_B}$ is the number-of-blocks information that the element was in the subset B. In view of the multiplicative nature of the number-of-blocks information, we need to average these block values over the blocks of a partition π using the multiplicative or geometric mean instead of the arithmetical mean. This can be done by considering the "associated number of blocks" random variable $H_{m,\pi}(u) = \frac{1}{p_B}$ if $u \in B$ which is the reciprocal of the partition indicator function $I_\pi(u) = p_B$ if $u \in B$. Then the geometric mean is formed by taking the product of these values over all $u \in U$ and then taking the $|U|^{th}$ root. This defines the *multiplicative (Shannon)*

entropy of a partition π (which does not involve any choice of a base for logs):

$$H_m(\pi) = \sqrt[|U|]{\prod_{u\in U} H_{m,\pi}(u)} = \sqrt[|U|]{\prod_{B\in\pi}\left(\frac{1}{p_B}\right)^{|B|}} = \prod_{B\in\pi}\left(\frac{1}{p_B}\right)^{p_B} \text{ blocks.}^{15}$$

It is useful to introduce the multiplicative or geometric expectation: given a finite-valued random variable X with the values $\{x_1, ..., x_n\}$ with the probabilities $\{p_1, ..., p_n\}$, the *multiplicative expectation* is $E_m[X] = \prod_{i=1}^{n} x_i^{p_i}$. Then the multiplicative entropy of a partition is the multiplicative expectation of the block values $H_m(B) = \frac{1}{p_B}$ which are the associated number of blocks. The usual (additive) Shannon entropy is then obtained as the $\log_2$ version of this "log-free" number-of-blocks measure:

$$\log_2(H_m(\pi)) = \log\left(\prod_{B\in\pi}\left(\frac{1}{p_B}\right)^{p_B}\right) = \sum_{B\in\pi}\log\left(\left(\frac{1}{p_B}\right)^{p_B}\right) = \sum_{B\in\pi} p_B \log\left(\frac{1}{p_B}\right) = H(\pi)$$

(which also justifies calling H_m the multiplicative *Shannon* entropy). Or viewed the other way around, $H_m(\pi) = 2^{H(\pi)}$. For the discrete partition on U, each p_B is $\frac{1}{|U|}$ so the multiplicative entropy of the discrete partition is $H_m\left(\widehat{0}\right) = \prod_{u\in U} |U|^{1/|U|} = |U|$ which could also be obtained as $2^{H\left(\widehat{0}\right)}$ since $H\left(\widehat{0}\right) = \log(|U|)$. Since the multiplicative Shannon entropy of a partition π of m equally probable blocks is $\prod_{B\in\pi} \sqrt[m]{m} = m$, the natural choice of unit for the multiplicative entropy is "blocks" (or geometric average of the associated number of blocks just as the "bits" of the additive Shannon measure is actually the arithmetical average number of bits from the block values). Since $H_m\left(\widehat{0}\right) = |U|$, the multiplicative Shannon entropy of the discrete partition on a 3^n element set is 3^n elements (= blocks). Hence the Shannon entropy with base 3 would be: $\log_3 H_m\left(\widehat{0}\right) = \log_3(3^n) = n$ trits as in Example 3 above. The multiplicative and logical entropies of the discrete partition or of any equal-blocked partition are the block value entropies for each equal block, and the block value relationship between the multiplicative Shannon entropy and the logical entropy in general is:

$$h(B) = 1 - \frac{1}{H_m(B)}$$

where $h(B) = 1 - p_B$ and $H_m(B) = 1/p_B$, and where $h(B)$ is the probability that the second draw (in drawing a random pair) is from a different block. Solving for the multiplicative entropy gives: $H_m(B) = \frac{1}{1-h(B)}$ which also has a simple interpretation. Since $h(B)$ is the probability that the second draw is in a different block than the first draw, $1 - h(B)$ is the probability of any one of the equiprobable blocks being drawn (in the hypothetical equal-blocked partition with all blocks like B) so $\frac{1}{1-h(B)} = H_m(B)$ is the number of blocks in that partition. After the block relationship $h(B) = 1 - \frac{1}{H_m(B)}$, the logical entropy $h(\pi) = \sum_{B\in\pi} p_B h(B)$ is obtained as the (additive) expectation of the block values $h(B) = 1 - p_B$ while the multiplicative entropy is obtained as the multiplicative expectation $H_m(\pi) = \prod_{B\in\pi} H_m(B)^{p_B}$ of the block values $H_m(B) = \frac{1}{p_B}$. Hence the relationship between the two averages can no longer be given by a simple functional relationship, and similarly for the relationship between the logical entropy $h(\pi)$ and the usual Shannon entropy $H(\pi) = \log_2(H_m(\pi))$.

[15] Another way to arrive at this formula is to use the relationship between a block probability p_B of π and the associated number of blocks $\frac{1}{p_B}$ in a long sequence of independent samples of the number-of-blocks random variable $H_{m,\pi}$ where $H_{m,\pi}(u) = 1/p_B$ if $u \in B$. In a long sequence of N samples of $H_{m,\pi}$, there would tend to be $p_B N$ instances of the number of blocks $1/p_B$ so the probable or typical sequence would tend to have the probability $\prod_{B\in\pi}(p_B)^{Np_B} = p_\pi^N$ as if an event with probability p_π had occurred N times where $p_\pi = \prod_{B\in\pi}(p_B)^{p_B}$. But this means that it is as if the number-of-blocks random variable had taken the value $\frac{1}{p_\pi} = \prod_{B\in\pi}\left(\frac{1}{p_B}\right)^{p_B} = H_m(\pi)$ each time with probability p_π. Thus $H_m(\pi)$ is the average associated number of blocks for the partition π (see the asymptotic equipartition property in information theory texts such as [4] or, for a simpler treatment, [3]).

3.3 Independent Partitions

The usual Shannon entropy block value $H(B) = \log_2\left(\frac{1}{p_B}\right)$ is the number of independent equal-blocked binary partitions that need to be intersected to create all the distinctions of the equal-blocked partition with $\frac{1}{p_B}$ blocks (as it were). What is the corresponding interpretation of the base-free multiplicative entropy block values $H_m(B) = \frac{1}{p_B}$? Since some of the blocks in the intersection of two partitions might be empty, the general number-of-blocks inequality is: $|\pi \wedge \sigma| \leq |\pi|\,|\sigma|$. When independent partitions intersect, the number of blocks multiplies so the inequality is an equation in that instance. Much verbiage has been expended arguing that some intuitive notion of "information" should be additive for independent partitions but the underlying mathematical fact is simply that the number of blocks is multiplicative for independent partitions and Shannon chose to use the logarithm of the number of blocks as his measure of information. When independent equal-blocked binary partitions are intersected, the number of blocks grows by the power of 2, e.g., $|\pi_1 \wedge \pi_2 \wedge ... \wedge \pi_m| = |\pi_1|\,|\pi_2|\,...\,|\pi_m| = 2^m$, so the logarithm of the number of blocks (e.g., $\log_2(2^m) = m$) just picks up the number of intersecting partitions. For the base-free number-of-blocks multiplicative notion of entropy, the question is how many independent equal-blocked partitions does it take to create all the dits of the hypothetical equal-blocked partition with $\frac{1}{p_B}$ blocks? The answer is 1, that partition itself which has $\frac{1}{p_B} = H_m(B)$ blocks. Hence that is the block count in the equal-blocked partition which creates those dits. Then taking the multiplicative expectation of those block counts $H_m(B)$ yields the multiplicative entropy $H_m(\pi)$ of the partition π. Moreover, if two partitions $\pi = \{B\}_{B\in\pi}$ and $\sigma = \{C\}_{C\in\sigma}$ are independent, then the block counts multiply, i.e., $H_m(B \cap C) = \frac{1}{p_{B\cap C}} = \frac{1}{p_B}\frac{1}{p_C} = H_m(B)\,H_m(C)$, so for the multiplicative expectations we have:

$$H_m(\pi \wedge \sigma) = \prod_{B,C} H_m(B \cap C)^{p_{B\cap C}} = \prod_{B,C} [H_m(B)H_m(C)]^{p_B p_C} = \left(\prod_{B\in\pi} H_m(B)^{p_B}\right)\left(\prod_{C\in\sigma} H_m(C)^{p_C}\right) = H_m(\pi)\,H_m(\sigma),$$

or taking logs,

$$H(\pi \wedge \sigma) = \log_2(H_m(\pi \wedge \sigma)) = \log_2(H_m(\pi)\,H_m(\sigma)) = \log_2(H_m(\pi)) + \log_2(H_m(\sigma)) = H(\pi) + H(\sigma).$$

Thus the Shannon entropy for the combined independent partitions (i.e., their "meet" in the usual lattice of partitions) is the sum of the entropies for the separate partitions, i.e., $H(\pi \wedge \sigma) = H(\pi) + H(\sigma)$. This intentional feature of the Shannon measure of information is motivated by the intuition that the "information" in two independent probability distributions is in some sense "disjoint." But we have seen that when the information in a partition is represented by its dit set $\mathrm{dit}(\pi)$, then the overlap in the dit sets of any two non-blob partitions is always non-empty. The dit set of the meet of two partitions is just the union, $\mathrm{dit}(\pi \wedge \sigma) = \mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma)$, so that union is never a disjoint union (when the dit sets are non-empty). The relationship of independence between partitions implies the same for their dit sets as events.

Proposition 3 *If π and σ are independent partitions, then their dit sets $\mathrm{dit}(\pi)$ and $\mathrm{dit}(\sigma)$ are independent as events in the sample space $U \times U$ (with equiprobable points).*

We need to show that the probability $m(\pi, \sigma)$ of the event $\mathrm{Mut}(\pi, \sigma) = \mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma)$ is equal to the product of the probabilities $h(\pi)$ and $h(\sigma)$ of the events $\mathrm{dit}(\pi)$ and $\mathrm{dit}(\sigma)$. By the assumption of independence, we have $\frac{|B\cap C|}{|U|} = p_{B\cap C} = p_B p_C = \frac{|B||C|}{|U|^2}$ so that $|B \cap C| = |B|\,|C|\,/\,|U|$. By the previous structure theorem for the mutual information set: $\mathrm{Mut}(\pi, \sigma) = \bigcup_{B\in\pi, C\in\sigma} (B - (B \cap C)) \times (C - (B \cap C))$, where the union is disjoint so that:

$$m(\pi,\sigma)=\frac{|\mathrm{Mut}(\pi,\sigma)|}{|U|^2}=\frac{1}{|U|^2}\sum_{B\in\pi,C\in\sigma}\left(|B|-|B\cap C|\right)\left(|C|-|B\cap C|\right)=$$
$$\frac{1}{|U|^2}\sum_{B\in\pi,C\in\sigma}\left(|B|-\frac{|B||C|}{|U|}\right)\left(|C|-\frac{|B||C|}{|U|}\right)$$
$$=\sum_{B\in\pi,C\in\sigma}\frac{|B|}{|U|}\left(1-\frac{|C|}{|U|}\right)\frac{|C|}{|U|}\left(1-\frac{|B|}{|U|}\right)=\sum_{B\in\pi,C\in\sigma}p_B\left(1-p_B\right)p_C\left(1-p_C\right)=$$
$$\sum_{B\in\pi}p_B\left(1-p_B\right)\sum_{C\in\sigma}p_C\left(1-p_C\right)=h\left(\pi\right)h\left(\sigma\right).$$

Hence under independence, the normalized dit count $m(\pi,\sigma)=\left|\mathrm{Mut}\left(\pi,\sigma\right)\right|/\left|U\right|^2$ of the mutual information set $\mathrm{Mut}\left(\pi,\sigma\right)=\mathrm{dit}\left(\pi\right)\cap\mathrm{dit}\left(\sigma\right)$ is equal to product of the normalized dit counts of the partitions:

$$\frac{|\mathrm{Mut}(\pi,\sigma)|}{|U\times U|}=m(\pi,\sigma)=h\left(\pi\right)h\left(\sigma\right)=\frac{|\mathrm{dit}(\pi)|}{|U|}\frac{|\mathrm{dit}(\sigma)|}{|U|}\text{ if }\pi\text{ and }\sigma\text{ are independent.}$$

This comparison of independence for the two entropy notions may misleadingly suggests that Shannon's entropy and logical entropy might be related as "additive" is related to "multiplicative." But many of the other corresponding formulas for the two entropy notions have the same form, e.g., are both additive. And we saw that there is already a multiplicative version $H_m\left(\pi\right)$ of Shannon's entropy measure.

3.4 Some Concepts for Shannon and Logical Entropies

3.4.1 Conditional Entropy and Mutual Information

For each block B in a partition π, the Shannon block value $H\left(B\right)=\log_2(\frac{1}{p_B})$ counts the number of independent equal-blocked binary partitions that need to be intersected to create the dits in the hypothetical equal-blocked partition with $\frac{1}{p_B}$ blocks (as it were) while the multiplicative block value $H_m(B)=\frac{1}{p_B}$ just counts the number of blocks in that equal-blocked partition with those same dits. The logical entropy block value $h\left(B\right)=1-p_B$ underlies both concepts since it is simply the number of dits itself (normalized) in that equal-blocked partition with $\frac{1}{p_B}=\frac{|U|}{|B|}$ blocks, i.e., the number of dits $\frac{|U|}{|B|}\times\left(|B|\,|U-B|\right)$ normalizes to $\frac{1}{|U|^2}\times\frac{|U|}{|B|}\times\left(|B|\,|U-B|\right)=\frac{|U-B|}{|U|}=1-\frac{|B|}{|U|}=1-p_B=h\left(B\right)$. Hence for each of the major concepts in the information theory based on the usual Shannon measure, there should be a corresponding concept based on the normalized dit counts of logical entropy. In the following sections, we give some of these analogous concepts and analogous results.

The definition of conditional entropy in conventional information theory is based on subset reasoning which is then averaged over a partition. Given a subset C which is a block in a partition σ, a partition $\pi=\{B\}_{B\in\pi}$ induces a partition of C with the blocks $\{B\cap C\}_{B\in\pi}$. Then $\left\{p_{B|C}\right\}_{B\in\pi}$ with $p_{B|C}=\frac{p_{B\cap C}}{p_C}$ is the probability distribution associated with that partition so it has an entropy which we denote: $H\left(\pi|C\right)=\sum_{B\in\pi}p_{B|C}\log\left(\frac{1}{p_{B|C}}\right)=\sum_B\frac{p_{B\cap C}}{p_C}\log\left(\frac{p_C}{p_{B\cap C}}\right)$. The (Shannon) *conditional entropy* is then obtained by averaging over the blocks of σ:

$$H\left(\pi|\sigma\right)=\sum_{C\in\sigma}p_CH\left(\pi|C\right)=\sum_{C\in\sigma}p_C\sum_B\frac{p_{B\cap C}}{p_C}\log\left(\frac{p_C}{p_{B\cap C}}\right)=\sum_{B,C}p_{B\cap C}\log\left(\frac{p_C}{p_{B\cap C}}\right)$$
$$=\sum_C\left[p_C\log\left(p_C\right)-\sum_Bp_{B\cap C}\log\left(p_{B\cap C}\right)\right]=H\left(\pi\wedge\sigma\right)-H\left(\sigma\right).$$

Since $H\left(\pi\wedge\sigma\right)=H\left(\pi|\sigma\right)+H\left(\sigma\right)$, the conditional entropy $H\left(\pi|\sigma\right)$ is thought of as the information contained in π that is not already in σ, i.e., the information that π adds onto the information in σ. If one considered an analogy with a Venn diagram with two circles $H\left(\pi\right)$ and $H\left(\sigma\right)$, then $H\left(\pi\wedge\sigma\right)$ would correspond to the union of the two circles and $H\left(\pi|\sigma\right)$ would correspond to the union after $H\left(\sigma\right)$ was subtracted out, i.e., $H\left(\pi\wedge\sigma\right)-H\left(\sigma\right)$.

The Venn diagram analogy raises the concept of the intersection of the two circles as a concept of "mutual information" in (conventional) information theory. We might apply the Venn diagram heuristics using a block $B \in \pi$ and a block $C \in \sigma$. We saw before that the information contained in a block B was $H(B) = \log\left(\frac{1}{p_B}\right)$ and similarly for C while $H(B \cap C) = \log\left(\frac{1}{p_{B\cap C}}\right)$ would correspond to the union of the information in B and in C. Hence the overlap or "mutual information" in B and C could be obtained as the sum of the two informations minus the union:

$$I(B;C) = \log\left(\frac{1}{p_B}\right) + \log\left(\frac{1}{p_C}\right) - \log\left(\frac{1}{p_{B\cap C}}\right) = \log\left(\frac{1}{p_B p_C}\right) + \log\left(p_{B\cap C}\right) = \log\left(\frac{p_{B\cap C}}{p_B p_C}\right).$$

Then the (Shannon) *mutual information* in the two partitions is obtained by averaging over the mutual information for each pair of blocks from the two partitions:

$$I(\pi;\sigma) = \sum_{B,C} p_{B\cap C} \log\left(\frac{p_{B\cap C}}{p_B p_C}\right).$$

The mutual information can be expanded to verify the Venn diagram heuristics:

$$I(\pi;\sigma) = \sum_{B,C} p_{B\cap C} \log\left(\frac{p_{B\cap C}}{p_B p_C}\right) =$$
$$\sum_{B,C} p_{B\cap C} \log\left(p_{B\cap C}\right) + \sum_{B,C} p_{B\cap C} \log\left(\frac{1}{p_B}\right) + \sum_{B,C} p_{B\cap C} \log\left(\frac{1}{p_C}\right)$$
$$= -H(\pi \wedge \sigma) + \sum_B p_B \log\left(\frac{1}{p_B}\right) + \sum_C p_C \log\left(\frac{1}{p_C}\right) = H(\pi) + H(\sigma) - H(\pi \wedge \sigma).$$

In the logical theory, the definitions of these concepts of mutual information and conditional entropy are simple and direct. Since the information in a partition π is given in its dit set $\mathrm{dit}(\pi)$, the information common to two partitions is the set of dits common to the two dit sets which normalizes to:

$$\textit{mutual logical information}\text{: } m(\pi,\sigma) = \frac{|\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)|}{|U|^2}.$$

Since the conditional entropy of a partition π given σ is the extra information in π not present in σ, it is given by the difference between their dit sets which normalizes to:

$$\textit{conditional logical entropy}\text{: } h(\pi|\sigma) = \frac{|\mathrm{dit}(\pi)-\mathrm{dit}(\sigma)|}{|U|^2}.$$

Since these notions are defined as the normalized size of subsets of the set of ordered pairs U^2, the Venn diagrams and inclusion-exclusion principle are not just analogies. For instance $|\mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma)| = |\mathrm{dit}(\pi)| + |\mathrm{dit}(\sigma)| - |\mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma)|$ so normalizing yields a version of the "modular law":

$$m(\pi,\sigma) = \frac{|\mathrm{dit}(\pi)\cap\mathrm{dit}(\sigma)|}{|U|^2} = \frac{|\mathrm{dit}(\pi)|}{|U|^2} + \frac{|\mathrm{dit}(\sigma)|}{|U|^2} - \frac{|\mathrm{dit}(\pi)\cup\mathrm{dit}(\sigma)|}{|U|^2} = h(\pi) + h(\sigma) - h(\pi \wedge \sigma).$$

In a similar manner, $|\mathrm{dit}(\pi) - \mathrm{dit}(\sigma)| = |\mathrm{dit}(\pi)| - |\mathrm{dit}(\pi) \cap \mathrm{dit}(\sigma)| = |\mathrm{dit}(\pi) \cup \mathrm{dit}(\sigma)| - |\mathrm{dit}(\sigma)|$ so normalizing yields:

$$h(\pi|\sigma) = h(\pi) - m(\pi,\sigma) = h(\pi \wedge \sigma) - h(\sigma).$$

Since the formulas in the two cases often have similar relationships, e.g., $H(\pi|\sigma) = H(\pi \wedge \sigma) - H(\sigma)$ and $h(\pi|\sigma) = h(\pi \wedge \sigma) - h(\sigma)$, it is useful to also emphasize some crucial differences. One of the most important special cases is the behavior of the two types of entropy for two partitions that are (stochastically) independent. Recall that partitions $\pi = \{B\}_{B\in\pi}$ and $\sigma = \{C\}_{C\in\sigma}$ are (stochastically) *independent* if for any $B \in \pi$ and $C \in \sigma$, $p_{B\cap C} = p_B p_C$. For independent partitions, it is immediate that $I(\pi;\sigma) = \sum_{B,C} p_{B\cap C} \log\left(\frac{p_{B\cap C}}{p_B p_C}\right) = 0$ but we have already seen that for the logical mutual information, $m(\pi,\sigma) > 0$ so long as neither partition is the blob $\widehat{1}$. However for independent partitions we have;

$$m(\pi,\sigma) = h(\pi)\, h(\sigma)$$

so the logical mutual information behaves like the probability of both events occurring in the case of independence. Indeed, the relation $m(\pi,\sigma) = h(\pi)\, h(\sigma)$ means that the probability that a randomly chosen pair (always with replacement) is distinguished by both partitions is the same as the probability that it is distinguished by one partition times the probability that it is distinguished by the other partition.

For Shannon's conditional entropy, independence implies that $H(\pi|\sigma) = H(\pi) - I(\pi;\sigma) = H(\pi)$ so in this case, it is Shannon's concept that behaves like the corresponding probability concept of conditional probability. For the logical conditional entropy, independence implies that $h(\pi|\sigma) = h(\pi) - m(\pi,\sigma) = h(\pi) - h(\pi)\, h(\sigma) = h(\pi)\,[1 - h(\sigma)]$. This means that if π and σ are independent, then the probability that a randomly chosen pair is distinguished by π but not by σ is the probability that it was distinguished by π times the probability that it was not distinguished by σ.

Independence has an even simpler probabilistic interpretation. In general,

$$[1 - h(\pi)]\,[1 - h(\sigma)] = 1 - h(\pi) - h(\sigma) + h(\pi)\, h(\sigma) = [1 - h(\pi \wedge \sigma)] + [h(\pi)\, h(\sigma) - m(\pi,\sigma]$$

which could also be rewritten as:

$$E[I_{\pi\wedge\sigma}] - E[I_\pi]\, E[I_\sigma] = [1 - h(\pi \wedge \sigma)] - [1 - h(\pi)]\,[1 - h(\sigma)] = m(\pi,\sigma) - h(\pi)\, h(\sigma)$$

so that:

$$\text{if } \pi \text{ and } \sigma \text{ are independent: } [1 - h(\pi)]\,[1 - h(\sigma)] = [1 - h(\pi \wedge \sigma)].$$

Thus if π and σ are independent, then the probability that a randomly drawn pair is identified by π *and* by σ is equal to the probability that the pair is identified by $\pi \wedge \sigma$.

3.4.2 Product Entropies and Co-Information

The *(logical) product entropy* (on U) is defined as:

$$h_U(\pi \times \sigma) = E[1 - I_\pi I_\sigma] = 1 - \textstyle\sum_{B,C} p_{B\cap C} p_B p_C.$$

This definition and the name can be directly motivated. For the partition $\pi = \{B\}_{B\in\pi}$ as a set of blocks, let $f_\pi : U \to \pi$ be the corresponding epimorphism where $f_\pi(u) = B$ {note "B" is here a point in the set π} if $u \in B$. Then from two partitions π and σ, the two epimorphisms f_π and f_σ induce an epimorphism $f_\pi \times f_\sigma = f_{\pi\times\sigma} : U \times U \to \pi \times \sigma$ which gives the *product partition* $\pi \times \sigma$ on $U \times U$ whose blocks are $\{B \times C\}_{B,C}$. Note that $U \times U$ is here used as the underlying set of a partition in addition to being the set of ordered pairs used to analyze partitions on U— two interpretations of $U \times U$ that we will exploit. The probability that a random element drawn from the underlying set $U \times U$ is in the block $B \times C$ is $p_B p_C$ so we have the partition indicator function: $I_{\pi\times\sigma} : U \times U \to [0,1]$ where $I_{\pi\times\sigma}(u,u') = p_B p_C$ if $u \in B$ and $u' \in C$. This is also the probability that the first element drawn is from B and that the second element drawn independently (i.e., with replacement) is in C, which is not necessarily the same as the probability that a random pair drawn with replacement has both elements in $B \cap C$ (unless the partitions are independent). Indeed, $p_B p_C$ would be positive even when $B \cap C = \varnothing$ so that $p_{B\cap C} = 0$. If $I_{\pi\times\sigma}$ is preceded by the diagonal map $\Delta_U : U \to U \times U$, then we have the product $I_\pi I_\sigma$ of partition indicators, i.e., $I_\pi I_\sigma = I_{\pi\times\sigma}\Delta : U \xrightarrow{\Delta} U \times U \xrightarrow{I_{\pi\times\sigma}} [0,1]$. This motivates designating $E[1 - I_\pi I_\sigma]$ as the logical product entropy $h_U(\pi \times \sigma)$ on U. We have written the product entropy with a subscript as $h_U(\pi \times \sigma)$ to distinguish it from the related but different ordinary logical entropy of the product partition $\pi \times \sigma$ on the set $U \times U$, i.e., from

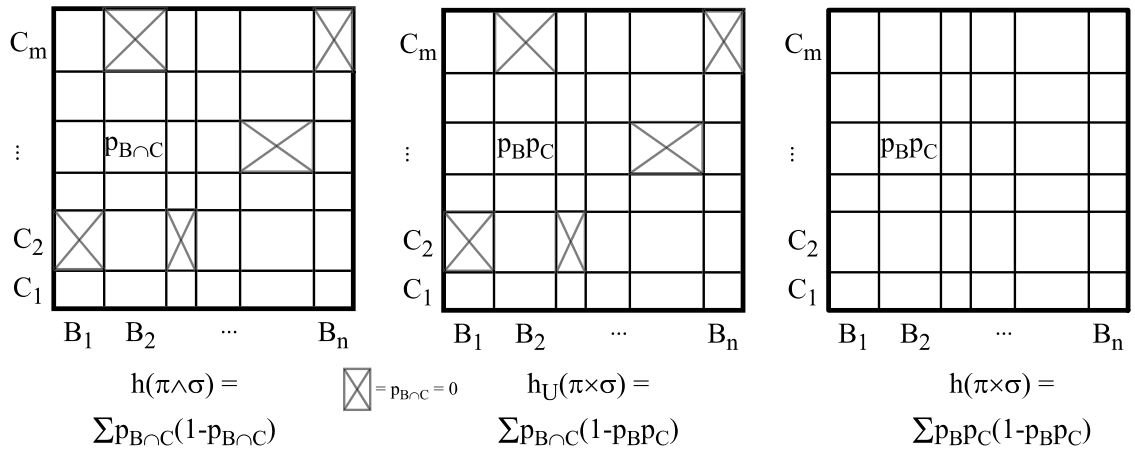
$$h(\pi \times \sigma) = \textstyle\sum_{B,C} p_B p_C\,(1 - p_B p_C) = 1 - \sum_{B,C} (p_B p_C)^2 = 1 - E[I_{\pi\times\sigma}] = 1 - E[I_\pi]\, E[I_\sigma] = 1 - [1 - h(\pi)]\,[1 - h(\sigma)].$$

Multiplying out $[1-h(\pi)][1-h(\sigma)]$ yields another modular law: $h(\pi\times\sigma) = h(\pi)+h(\sigma)-h(\pi)h(\sigma)$. Since $1-h(\pi\times\sigma) = [1-h(\pi)][1-h(\sigma)]$ (i.e., $E[I_{\pi\times\sigma}] = E[I_\pi]E[I_\sigma]$), there is a simple probabilistic interpretation that entwines a pair of ordered pairs on U and a pair of elements of $U\times U$ which is also a pair of ordered pairs. This is a "four draw" interpretation where we consider four independent random draws from U or, equivalently, two independent random draws of elements drawn from $U\times U$, say, (u_1,u_2) and (u_3,u_4). Then the probability that the pair is identified by $\pi\times\sigma$, i.e., $1-h(\pi\times\sigma)$, is equal to the probability that the first pair (u_1,u_2) is identified by π and the independent second pair (u_3,u_4) is identified by σ, i.e., $[1-h(\pi)][1-h(\sigma)]$.

The product entropy $h_U(\pi\times\sigma)$ on U has a "three-draw" interpretation using three independent drawings from U. The first element is in some unique set $B\cap C$ for $B\in\pi$ and $C\in\sigma$ so we consider the event that the second draw is in B and the third draw is in C. The probability of that the second draw is in the B and the third draw is in the C with both B and C determined by the first draw is: $E[I_\pi I_\sigma] = \sum_{B,C} p_{B\cap C}p_Bp_C$. Hence $h_U(\pi\times\sigma) = E[1-I_\pi I_\sigma] = 1-E[I_\pi I_\sigma]$ is the probability that either the second draw does not fall in the B or the third draw does not fall in the C, where the B and C are determined by the first draw. This might be compared to the probabilistic interpretation of the meet entropy $h(\pi\wedge\sigma)$ which is that in two independent draws, the second draw does not fall in the same $B\cap C$ as the first draw.

The dit set of the "meet" $\pi\wedge\sigma$ is the *union* of the dit sets of the partitions, i.e., $\operatorname{dit}(\pi\wedge\sigma) = \operatorname{dit}(\pi)\cup\operatorname{dit}(\sigma)$, and the modular law for dit sets: $\operatorname{dit}(\pi)\cup\operatorname{dit}(\sigma) = \operatorname{dit}(\pi)+\operatorname{dit}(\sigma)-\operatorname{dit}(\pi)\cap\operatorname{dit}\sigma$ normalizes to: $h(\pi\wedge\sigma) = h(\pi)+h(\sigma)-m(\pi,\sigma)$.[16] The meet entropy and the two entropies $h_U(\pi\times\sigma)$ and $h(\pi\times\sigma)$ are all closely related to the union of information in the two partitions. If the two partitions are independent, then the following three union-related entropies are equal:

$$h(\pi\wedge\sigma) = \sum_{B,C} p_{B\cap C}(1-p_{B\cap C}) = 1-E[I_{\pi\wedge\sigma}] \text{ (entropy of meet)}$$
$$h_U(\pi\times\sigma) = \sum_{B,C} p_{B\cap C}(1-p_Bp_C) = 1-E[I_\pi I_\sigma] \text{ (product entropy on } U),$$
$$h(\pi\times\sigma) = \sum_{B,C} p_Bp_C(1-p_Bp_C) = 1-E[I_\pi]E[I_\sigma] \text{ (entropy of product on } U\times U).$$



Three Union-Related Entropies

The relationship of the three union-related entropies to each other is given by the inequality:

$$0 \le 2h_U(\pi\times\sigma) - h(\pi\wedge\sigma) - h(\pi\times\sigma) \text{ with equality under independence.}$$

which can be derived by a simple sum-of-squares calculation.

[16] Recall the earlier complaint that the usual order in the lattice of partitions should be written the other way around, i.e., according to the inclusion of dit sets rather than indit sets. Then the "meet" of partitions would become their join in accordance with the union of dit sets.

$$0 \leq \sum_{B,C} \left(p_{B\cap C} - p_B p_C\right)^2 = \sum_{B,C} p_{B\cap C}^2 - 2\sum_{B,C} p_{B\cap C} p_B p_C + \sum_{B,C} \left(p_B p_C\right)^2$$
$$= \left[1 - h\left(\pi \wedge \sigma\right)\right] - 2\left[1 - h_U\left(\pi \times \sigma\right)\right] + \left[1 - h\left(\pi \times \sigma\right)\right] = 2h_U\left(\pi \times \sigma\right) - h(\pi \wedge \sigma) - h\left(\pi \times \sigma\right).$$

This inequality written in the form $h(\pi \wedge \sigma) \leq 2h_U\left(\pi \times \sigma\right) - h\left(\pi \times \sigma\right)$ might be taken as the logical entropy counterpart to the Shannon entropy inequality $H\left(\pi \wedge \sigma\right) \leq H\left(\pi\right) + H\left(\sigma\right)$ or, multiplicatively, $H_m(\pi \wedge \sigma) \leq H_m\left(\pi\right) H_m\left(\sigma\right)$ both of which have equality under independence of the two partitions.

The union-related product entropy $h_U\left(\pi \times \sigma\right)$ suggests another "intersection-related" concept that would be related to the mutual information $m(\pi, \sigma)$ (normalized intersection of dit sets) as the product entropy $h_U\left(\pi \times \sigma\right)$ if related to the entropy $h\left(\pi \wedge \sigma\right)$ of the meet. Instead of considering the probability $E\left[I_\pi I_\sigma\right]$ that the second and third draws are respectively in the B and C determined by the first draw, consider the probability $E\left[\left(1 - I_\pi\right)\left(1 - I_\sigma\right)\right]$ that the second draw was not in B *and* that the third draw was not in C where the B and C were determined by the first draw being in $B \cap C$. This quantity might be called the *logical co-information* and denoted:

$$c\left(\pi, \sigma\right) = E\left[\left(1 - I_\pi\right)\left(1 - I_\sigma\right)\right] = E\left[1 - I_\pi - I_\sigma + I_\pi I_\sigma\right] = E\left[\left(1 - I_\pi\right) + \left(1 - I_\sigma\right) - \left(1 - I_\pi I_\sigma\right)\right]$$
$$= E\left[1 - I_\pi\right] + E\left[1 - I_\sigma\right] - E\left[1 - I_\pi I_\sigma\right] = h\left(\pi\right) + h\left(\sigma\right) - h_U\left(\pi \times \sigma\right).$$

Thus we have the "modular law": $c\left(\pi, \sigma\right) = h\left(\pi\right) + h\left(\sigma\right) - h_U\left(\pi \times \sigma\right)$ holding where $c\left(\pi, \sigma\right)$ is analogous to $m\left(\pi, \sigma\right)$ and $h_U\left(\pi \times \sigma\right)$ is analogous to $h\left(\pi \wedge \sigma\right)$ in the previous "modular law": $m\left(\pi, \sigma\right) = h\left(\pi\right) + h\left(\sigma\right) - h\left(\pi \wedge \sigma\right)$ which was based on the normalized sizes of the dit sets. If π and σ are independent, then the two modular laws become the same since $h_U\left(\pi \times \sigma\right) = h\left(\pi \wedge \sigma\right)$ and thus $c\left(\pi, \sigma\right) = m\left(\pi, \sigma\right) = h\left(\pi\right) h\left(\sigma\right)$.

In summary, there are three union-related quantities that are all equal under independence:

$$h\left(\pi \wedge \sigma\right) = 1 - E\left[I_{\pi\wedge\sigma}\right]$$
$$h_U\left(\pi \times \sigma\right) = 1 - E\left[I_\pi I_\sigma\right]$$
$$h\left(\pi \times \sigma\right) = 1 - E\left[I_\pi\right] E\left[I_\sigma\right].$$

There are three intersection-related quantities that are all equal under independence:

$$m\left(\pi, \sigma\right) = E\left[1 - I_\pi - I_\sigma + I_{\pi\wedge\sigma}\right]$$
$$c\left(\pi, \sigma\right) = E\left[\left(1 - I_\pi\right)\left(1 - I_\sigma\right)\right]$$
$$h\left(\pi\right) h\left(\sigma\right) = E\left[1 - I_\pi\right] E\left[1 - I_\sigma\right].$$

And there are the corresponding three modular laws;

$$h\left(\pi \wedge \sigma\right) = h\left(\pi\right) + h\left(\sigma\right) - m\left(\pi, \sigma\right)$$
$$h_U\left(\pi \times \sigma\right) = h\left(\pi\right) + h\left(\sigma\right) - c\left(\pi, \sigma\right)$$
$$h\left(\pi \times \sigma\right) = h\left(\pi\right) + h\left(\sigma\right) - h\left(\pi\right) h\left(\sigma\right).$$

Thus, in general (i.e., not just under independence), we have the equations:

$$h\left(\pi\right) + h\left(\sigma\right) = h\left(\pi \wedge \sigma\right) + m\left(\pi, \sigma\right) = h_U\left(\pi \times \sigma\right) + c\left(\pi, \sigma\right) = h\left(\pi \times \sigma\right) + h\left(\pi\right) h\left(\sigma\right).$$

The union-related entropies are additionally related by the inequality:

$$0 \leq 2h_U\left(\pi \times \sigma\right) - h(\pi \wedge \sigma) - h\left(\pi \times \sigma\right) \text{ with equality under independence}$$

which implies the following inequality between the intersection-related entropies:

$$0 \leq m\left(\pi, \sigma\right) + h\left(\pi\right) h\left(\sigma\right) - 2c\left(\pi, \sigma\right) \text{ with equality under independence.}$$

Shannon's mutual information $I(\pi;\sigma) = \sum_{B,C} p_{B\cap C} \log\left(\frac{p_{B\cap C}}{p_B p_C}\right)$ satisfies the inequality: $0 \leq I(\pi;\sigma)$ with equality under independence. One of the principal design features of Shannon's measure was that it gave the two inequalities: $0 \leq I(\pi;\sigma)$ and $H(\pi \wedge \sigma) \leq H(\pi) + H(\sigma)$ with equality under independence. These entropies also satisfy the Venn diagram heuristics: $I(\pi;\sigma) = H(\pi) + H(\sigma) - H(\pi \wedge \sigma)$. We noted previously that when information is measured by the normalized counts of distinctions, then the mutual information $m(\pi,\sigma)$ in the basic sense of the distinctions common to the two partitions is always strictly positive for non-blob partitions. The Venn diagram for dit sets gives us the corresponding modular law: $m(\pi,\sigma) = h(\pi) + h(\sigma) - h(\pi \wedge \sigma)$, and thus we also have the inequalities: $0 \leq m(\pi,\sigma)$ and $h(\pi \wedge \sigma) \leq h(\pi) + h(\sigma)$. These inequalities are strict for non-blob partitions but they are not equations when the partitions are independent. Hence the question arises of what are the inequalities in the logical case that would be analogous to the two Shannon inequalities above which are equations under independence. The analogue to $0 \leq I(\pi;\sigma)$ comes from the other intersection-related concepts in addition to $m(\pi,\sigma)$, namely the inequality: $0 \leq m(\pi,\sigma) + h(\pi)\, h(\sigma) - 2c(\pi,\sigma)$ with equality under independence. And the analogue to the inequality $H(\pi \wedge \sigma) \leq H(\pi) + H(\sigma)$ comes from the other union-related concepts in addition to $h(\pi \wedge \sigma)$, namely the inequality: $h(\pi \wedge \sigma) \leq 2h_U(\pi \times \sigma) - h(\pi \times \sigma)$ with equality under independence.

3.4.3 Cross Entropy and Divergence

Given a subset $S \subseteq U$ for finite U, probability theory started with the "natural" or Laplacian definition of the probability of the event S as its normalized size $p_S = \frac{|S|}{|U|}$, and then the theory was generalized to deal with other probability distributions on the subsets of a finite sample space. Given a set partition $\pi = \{B\}_{B\in\pi}$ on a set U, we have used the "natural" or Laplacian probability distribution $p_B = \frac{|B|}{|U|}$ determined by the partition. The set partition π also determines the set of distinctions $\operatorname{dit}(\pi) \subseteq U \times U$ and the logical entropy of the partition is the Laplacian probability of the dit set as an event, i.e., $h(\pi) = \frac{|\operatorname{dit}(\pi)|}{|U\times U|} = \sum_B p_B(1 - p_B)$. But we may also "kick away the ladder" and generalize all the definitions to any finite probability distributions $p = \{p_1, ..., p_n\}$. A probability distribution p might be given by finite-valued random variables X on a sample space U where $p_i = \operatorname{Prob}(X = x_i)$ for the finite set of distinct values x_i for $i = 1, ..., n$. Thus the logical entropy of the random variable X is: $h(X) = \sum_{i=1}^n p_i(1 - p_i) = 1 - \sum_i p_i^2$. The entropy is only a function of the probability distribution of the random variable, not its values, so we could also take it simply as a function of the probability distribution p, $h(p) = 1 - \sum_i p_i^2$. Taking the sample space as $\{1, ..., n\}$, the logical entropy is still interpreted as the probability that two independent samples draw distinct points from $\{1, ..., n\}$. The further generalizations replacing probabilities by probability density functions and sums by integrals are clear but beyond the scope of this introduction.

Once we cut the probability distribution defined on the blocks of a partition loose from the Laplacian definition, we can consider multiple distributions at the same time. For instance, on the partition $\pi \wedge \sigma = \{B \cap C\}_{B\in\pi, C\in\sigma}$, we have one distribution defined on the blocks which is the Laplacian distribution $p_{B\cap C} = \frac{|B\cap C|}{|U|}$ but we also have the distribution which assigns $p_B p_C$ to the block $B \cap C$. Then we can consider draws that select blocks according to a probability distribution on the blocks (which is only the same as selecting the block according to a randomly selected point in U for the Laplacian distribution). Then the probability of selecting distinct blocks when the first block is selected according to the distribution $\{p_{B\cap C}\}$ and the second block is selected according to $\{p_B p_C\}$ is: $\sum_{B,C} p_{B\cap C}(1 - p_B p_C) = h_U(\pi \times \sigma)$, the product entropy on U (this is different from the three-draw interpretation of the product entropy previously given). This example motivates the following definition of "cross entropy."

Given two probability distributions $p = \{p_1, ..., p_n\}$ and $q = \{q_1, ..., q_n\}$ on the same sample space $\{1, ..., n\}$, their *cross entropy* is defined as:

$$h(p\|q) = \sum_i p_i(1 - q_i) = 1 - \sum_i p_i q_i = \sum_i q_i(1 - p_i) = h(q\|p)$$

which is symmetric. Picking one point from $\{1,...,n\}$ according to p and another point according to q, the cross entropy $h(p\|q)$ is the probability that the points are distinct.

The notion of cross entropy in Shannon's information theory is: $H(p\|q) = \sum_i p_i \log\left(\frac{1}{q_i}\right)$ which is not symmetrical due to the asymmetric role of the logarithm. Then the *Kullback-Leibler divergence* $D(p\|q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$ is defined as a measure of the distance or divergence between the two distributions where $D(p\|q) = H(p\|q) - H(p)$. The "information inequality" is that: $D(p\|q) \geq 0$ with equality if and only if $p_i = q_i$ for $i = 1,...,n$ [4, p. 26]. But starting afresh, one might ask: "What is the natural measure of the difference between two probability distributions $p = \{p_1,...,p_n\}$ and $q = \{q_1,...,q_n\}$ that would always be non-negative and zero if and only they are equal?" The answer is clearly the sum of differences squared—which we take as the definition of the *logical divergence* (or *relative entropy*): $d(p\|q) = \sum_i (p_i - q_i)^2$, which is obviously also symmetric.[17] We have component-wise:

$$0 \leq (p_i - q_i)^2 = p_i^2 - 2p_iq_i + q_i^2 = 2\left[\tfrac{1}{n} - p_iq_i\right] - \left[\tfrac{1}{n} - p_i^2\right] - \left[\tfrac{1}{n} - q_i^2\right]$$

so that taking the sum for $i = 1,...,n$ gives:

$$0 \leq \sum_i (p_i - q_i)^2 = \sum_i p_i^2 - 2\sum_i p_iq_i + \sum_i q_i^2 = 2\left[1 - \sum_i p_iq_i\right] - \left[1 - \sum_i p_i^2\right] - \left[1 - \sum_i q_i^2\right] = 2h(p\|q) - h(p) - h(q).$$

Thus the logical version of the information inequality becomes:

$$0 \leq d(p\|q) = 2h(p\|q) - h(p) - h(q) \text{ with equality if and only if } p_i = q_i \text{ for } i = 1,...,n.$$

The p-weighted differences with q are $\sum_i p_i(p_i - q_i)$ and added to the logical entropy of p give the cross entropy so we have:

$$h(p\|q) = h(p) + \sum_i p_i(p_i - q_i) = h(q) + \sum_i q_i(q_i - p_i)$$

This allows us to rewrite the divergence in several ways including as the sum of the two average differences:

$$0 \leq d(p\|q) = 2h(p\|q) - h(p) - h(q) = h(p\|q) + \sum_i p_i(p_i - q_i) - h(q)$$
$$= h(p\|q) + \sum_i q_i(q_i - p_i) - h(p) = \sum_i p_i(p_i - q_i) + \sum_i q_i(q_i - p_i)$$

Taking the two distributions as $\{p_{B\cap C}\}_{B,C}$ and $\{p_Bp_C\}_{B,C}$ defined on the point set of all the pairs $\{(B,C) : B \in \pi, C \in \sigma\}$, then we have:

$$h(\{p_{B\cap C}\} \parallel \{p_Bp_C\}) = h_U(\pi \times \sigma),$$
$$h(\{p_{B\cap C}\}) = h(\pi \wedge \sigma),$$
$$h(\{p_Bp_C\}) = h(\pi \times \sigma),$$

so that the inequality $0 \leq d(p\|q) = 2h(p\|q) - h(p) - h(q)$ then becomes the previously derived:

$$0 \leq d(\{p_{B\cap C}\} \parallel \{p_Bp_C\}) = 2h_U(\pi \times \sigma) - h(\pi \wedge \sigma) - h(\pi \times \sigma) \text{ with equality under independence.}$$

Another important special case of the information inequality is when $p = \{p_1,...,p_n\}$ is the uniform distribution with all $p_i = \frac{1}{n}$. Then $h(p) = 1 - \frac{1}{n}$ where the probability that a random pair is distinguished (i.e., the random variable X with $\text{Prob}(X = x_i) = p_i$ has different values in two independent samples) takes the specific form of the probability $1 - \frac{1}{n}$ that the second draw gets a different value than the first. It may at first seem counterintuitive that the cross entropy is

[17] In other words, the most natural measure of the "distance" between two finite probabilities on a sample space with n points is just the distance between them as n-vectors (or in this case, the square of the distance).

$h(p\|q) = 1 - \frac{1}{n} = h(p)$ for any $q = \{q_1, ..., q_n\}$. But $h(p\|q)$ is the probability that the two points, say i and i', in the sample space $\{1, ..., n\}$ are distinct when one draw was according to p and the other according to q. Taking the first draw according to q, the probability that the second draw is distinct from whatever point was determined in the first draw is indeed $1 - \frac{1}{n}$. Then the divergence $d(p\|q) = 2h(p\|q) - h(p) - h(q) = \left(1 - \frac{1}{n}\right) - h(q)$ is a non-negative measure of how much the probability distribution q diverges from the uniform distribution (with the divergence being zero when q is the uniform distribution). It is simply the difference in the probability that a random pair will be distinguished by the uniform distribution and by q. Also this shows that among all probability distributions on $\{1, ..., n\}$, the uniform distribution has the maximum logical entropy.

3.4.4 Some Inequalities in Logical Information Theory

Since the partition indicator functions take values in the unit interval $[0, 1]$, all the complements such as $1 - I_\pi$ and all the products such as $I_\pi I_\sigma$ also take values in the unit interval, and thus so do their expectations. The random variable $1 - I_\pi - I_\sigma + I_{\pi \wedge \sigma}$ has the value at $u \in U$ of $1 - p_B - p_C + p_{B \cap C} = 1 - p_{B \cup C}$ where $u \in B \cap C$ so it also takes values in the unit interval and thus so does its expectation $E[1 - I_\pi - I_\sigma + I_{\pi \wedge \sigma}] = m(\pi, \sigma)$. This was also indicated by the probabilistic interpretations given to all the various logical entropies and logical informations. The largest logical entropy is for the discrete partition $h\left(\widehat{0}\right) = 1 - \frac{1}{|U|}$ (the probability that the second draw is different from the first) so for all the logical entropies of partitions, we have: $0 \leq h(\pi) < 1$ with equality if and only if $\pi = \widehat{1}$, the blob. For the mutual information and co-information of two partitions, we also have: $0 \leq m(\pi, \sigma) < 1$ and $0 \leq c(\pi, \sigma) < 1$ with equality in each case if and only if one (or both) of the partitions is the blob. For the logical product entropy: $0 \leq h_U(\pi \times \sigma) < 1$ with equality if and only if both partitions are the blob.

For any partition π with the n probabilities $\{p_1, ..., p_n\}$, we saw previously that:

$$h(\pi) \leq 1 - \frac{1}{n} \text{ with equality if and only if } p_1 = ... = p_n = \frac{1}{n}.$$

For the corresponding results in the Shannon's information theory, we might also apply the arithmetic-geometric mean inequality. For any list of n non-negative real numbers $x_1, ..., x_n$ we have:

$$\sqrt[n]{x_1 ... x_n} \leq \frac{1}{n} \sum x_i \text{ where they are equal if and only if } x_1 = ... = x_n.$$

The multiplicative entropy is a geometric mean:

$$H_m(\pi) = \prod_{B \in \pi} \underbrace{\left(\frac{1}{p_B}\right)^{1/|U|} ... \left(\frac{1}{p_B}\right)^{1/|U|}}_{|B|} = \sqrt[|U|]{\prod_{u \in U} H_m(u)}$$

where $H_m(u) = 1/p_B$ if $u \in B$. Then the inequality is:

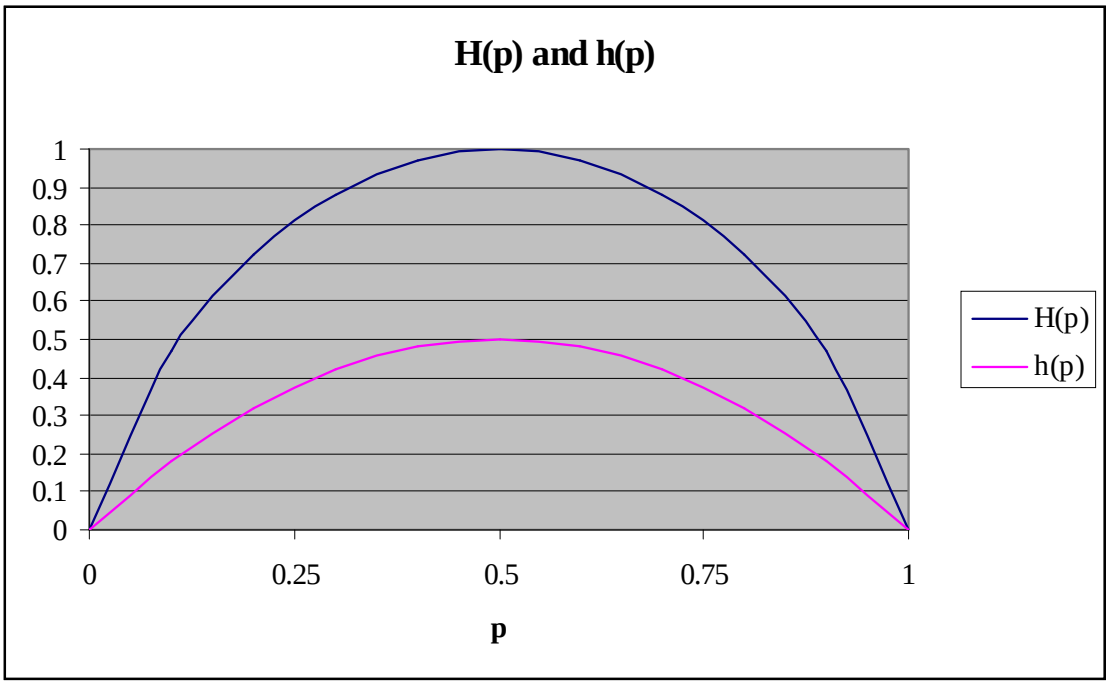
$$H_m(\pi) = \sqrt[|U|]{\prod_{u \in U} H_m(u)} \leq \frac{1}{|U|} \sum_u H_m(u) = \frac{1}{|U|} \sum_{B \in \pi} \frac{|B|}{p_B} = \sum_{B \in \pi} \frac{p_B}{p_B} = |\pi|, \text{ or}$$
$$H_m(\pi) \leq |\pi| \text{ with equality if and only if all the probabilities are equal, i.e., } p_1 = ... = p_n = \frac{1}{|\pi|}.$$

Taking logs we have:

$$H(\pi) \leq \log(|\pi|) \text{ with equality if and only if all the probabilities are equal}$$

so both the Shannon and logical entropies take their maximum values (for fixed number of blocks $|\pi|$) at the partition with equiprobable values.

In information theory texts, it is customary to graph the case of $n = 2$ where the entropy is graphed as a function of $p_1 = p$ with $p_2 = 1 - p$. The Shannon entropy function $H(p) = -p \log(p) - (1 - p) \log(1 - p)$ looks somewhat like an inverted parabola with its maximum value of $\log(n) = \log(2) = 1$ at $p = .5$. The logical entropy function $h(p) = 1 - p^2 - (1 - p)^2 = 2p - 2p^2 = 2p(1 - p)$ is an inverted parabola with its maximum value of $1 - \frac{1}{n} = 1 - \frac{1}{2} = .5$ at $p = .5$.



Given two n-tuples of reals, $(a_1, ..., a_n)$ and $(b_1, ..., b_n)$, then the *Cauchy-Schwartz inequality* is that:

$$\left(\sum_{i=1}^{n} a_i b_i\right)^2 \leq \left(\sum_i a_i^2\right)\left(\sum_i b_i^2\right)$$

with equality if and only if $\frac{a_1}{b_1} = ... = \frac{a_n}{b_n}$. For partitions π and σ, take the a_i's as $\{p_{B\cap C}\}_{B\in\pi, C\in\sigma}$ and the b_i's as $\{p_B p_C\}_{B\in\pi, C\in\sigma}$ so that the C-S inequality yields:

$$\begin{gathered}[1 - h_U(\pi \times \sigma)]^2 = E[I_\pi I_\sigma]^2 = \left(\textstyle\sum_{B,C} p_{B\cap C} p_B p_C\right)^2 \\ \leq \left(\textstyle\sum_{B,C} p_{B\cap C}^2\right)\left(\textstyle\sum_{B,C} p_B^2 p_C^2\right) = \left(\textstyle\sum_{B,C} p_{B\cap C}^2\right)\left(\textstyle\sum_B p_B^2\right)\left(\textstyle\sum_C p_C^2\right) = E[I_{\pi\wedge\sigma}] E[I_\pi] E[I_\sigma] \\ = [1 - h(\pi \wedge \sigma)][1 - h(\pi)][1 - h(\sigma)] \\ [1 - h(\pi \wedge \sigma)][1 - h(\pi \times \sigma)].\end{gathered}$$

In short,

$$[1 - h_U(\pi \times \sigma)]^2 \leq [1 - h(\pi \wedge \sigma)][1 - h(\pi \times \sigma)]$$

with equality if and only if $\frac{p_{B\cap C}}{p_B p_C} = k$ for some constant k and all $B \in \pi$ and $C \in \sigma$. If the equality holds, then $\sum_{B,C} p_{B\cap C} = 1 = k \sum_{B,C} p_B p_C = k$, so the equality holds if and only if π and σ are independent.

3.5 Summary of Analogous Concepts and Results

	Shannon Entropy	Logical Entropy
Block Entropy	$H(B) = \log(1/p_B)$	$h(B) = 1 - p_B$
Relationship	$H(B) = \log\left(\frac{1}{1-h(B)}\right)$	$h(B) = 1 - \frac{1}{2^{H(B)}}$
Entropy	$H(\pi) = \sum p_B \log(1/p_B)$	$h(\pi) = \sum p_B(1-p_B)$
Mutual Information	$I(\pi;\sigma) = H(\pi) + H(\sigma) - H(\pi \wedge \sigma)$	$m(\pi,\sigma) = h(\pi) + h(\sigma) - h(\pi \wedge \sigma)$
Independence	$I(\pi;\sigma) = 0$	$m(\pi,\sigma) = h(\pi)h(\sigma)$
Independence & Meets	$H(\pi \wedge \sigma) = H(\pi) + H(\sigma)$	$[1 - h(\pi \wedge \sigma)] = [1 - h(\pi)][1 - h(\sigma)]$
Conditional Entropy	$H(\pi\|\sigma) = H(\pi \wedge \sigma) - H(\sigma)$	$h(\pi\|\sigma) = h(\pi \wedge \sigma) - h(\sigma)$
Cross Entropy	$H(p\|\|q) = \sum p_i \log(1/q_i)$	$h(p\|\|q) = \sum p_i(1-q_i)$
Divergence	$D(p\|\|q) = H(p\|\|q) - H(p)$	$d(p\|\|q) = 2h(p\|\|q) - h(p) - h(q)$
Information Ineq.	$D(p\|\|q) \geq 0$ with $=$ iff $p_i = q_i \forall i$	$d(p\|\|q) \geq 0$ with $=$ iff $p_i = q_i \forall i$
Product Entropy	$H(\{p_{B\cap C}\} \,\|\|\, \{p_B p_C\})$	$h(\{p_{B\cap C}\} \,\|\|\, \{p_B p_C\}) = h_U(\pi \times \sigma)$
Entropy of Product	$H(\{p_B p_C\}) = \sum p_B p_C \log\left(\frac{1}{p_B p_C}\right)$	$h(\pi \times \sigma) = \sum p_B p_C (1 - p_B p_C)$
Info. Ineq. for Prod.	$I(\pi;\sigma) = D(\{p_{B\cap C}\} \,\|\|\, \{p_B p_C\}) \geq 0$ with equality under independence	$d(\{p_{B\cap C}\} \,\|\|\, \{p_B p_C\}) \geq 0$ with equality under independence.

3.6 The Noiseless Coding Theorem in Logical Information Theory

To build a bridge from logical information theory into coding theory, we develop and prove the logical information theory version of Shannon's Noiseless Coding Theorem [1, p. 73]. We started with the dual creation myth to visualize partitions as being created by adding distinctions to the blob. Given any rooted tree starting with the blob and ending with leaves corresponding to the elements of U, a partition on U would be determined by a maximal antichain of nodes in the tree, i.e., a set of nodes so that every leaf was the descendent of one and only one node. A node is then replaced by the set of its descendents in the "set of blocks" notion of a partition. Then we worked, for the most part, with the Laplacian probability assignment of relative block size.

Now we consider the reverse problem of starting with a finite probability distribution $P = \{p_1, ..., p_n\}$ and then developing rooted binary trees starting at the blob ("code trees" in coding theory) so that the probabilities in the given distribution and its powers are approximated arbitrarily well by the relative block sizes. For each p_i, let l_i be the least integer such that $\frac{1}{2^{l_i}} \leq p_i$. If $\frac{1}{2^{l_i}} \leq \frac{p_i}{2}$, then $\frac{1}{2^{l_i-1}} \leq p_i$ contrary to l_i being the least such integer so we have: $\frac{p_i}{2} < \frac{1}{2^{l_i}} \leq p_i$. Summing the inequality on the right, we have $\sum_i 2^{-l_i} \leq 1$. Let $l^* = \max\{l_i : i = 1, ..., n\}$. The complete binary tree (every node branches) with all the leaves at level l^* will have 2^{l^*} leaves which we could think of as the elements of a universe set U. A node at level l_i would have $2^{l^*-l_i}$ leaves as descendents so taking that set of leaves as a block in a partition, the relative block size is $\frac{2^{l^*-l_i}}{2^{l^*}} = 2^{-l_i}$. Thus the determination of the least l_i such that $\frac{1}{2^{l_i}} \leq p_i$ is the beginning of approximating the probability distribution $\{p_1, ..., p_n\}$ from below with the blocks of a partition generated from the blob by a binary tree.

How do we know that nodes with the probabilities 2^{-l_i} can always be selected to form an antichain (i.e., no two nodes related to each other by the partial ordering of the tree)? In coding theory, the *Kraft Inequality* (see [1, p. 57] or any other text for a proof) implies that the nodes with probabilities 2^{-l_i} can be selected to form an antichain in a binary tree if and only if $\sum_i 2^{-l_i} \leq 1$ (where nodes forming an antichain correspond to a prefix or instantaneous code assigned to the nodes). Since the Kraft condition $\sum_i 2^{-l_i} \leq 1$ is satisfied, we know that there is an antichain of n nodes in the complete binary tree with l^* levels and 2^{l^*} leaves so that the relative block sizes of the descendent leaves of those n nodes are 2^{-l_i} for $i = 1, ..., n$. But there is no reason for it to be a maximal antichain in general, i.e., $\sum_i 2^{-l_i} \leq 1$ might be a strict inequality.

Multiplying $\frac{p_i}{2} < 2^{-l_i} \leq p_i$ through by p_i and summing yields: $\frac{1}{2}\sum_i p_i^2 < \sum_i p_i 2^{-l_i} \leq \sum_i p_i^2$. When drawing random pairs from the universe set U, the sum $\sum_B p_B^2$ over the blocks $\{B\}_{B\in\pi}$ of

a partition π was the identification probability, i.e., the probability that the random pair would be identified by the partition. The sum $\sum_i p_i^2$ is the probability that a random pair picked from the set of outcomes $\{1, ..., n\}$ will be the same outcome, and $\sum_i p_i 2^{-l_i}$ is the probability that when a pair is drawn, the first draw according to the probabilities P and the second draw according to the probabilities $2^l = \left\{2^{-l_1}, ..., 2^{-l_n}, 1 - \sum_{i=1}^n 2^{-l_i}\right\}$, will be the same outcome (the extra outcome added to soak up any surplus probability $1 - \sum 2^{-l_i}$ has zero probability under P so it can be ignored). The complements of those identification probabilities are the distinction probabilities or entropies: $1 - \sum_i p_i^2 = h(P)$ and $1 - \sum_i p_i 2^{-l_i} = h\left(P \| 2^l\right)$. Hence the inequality can be written as:

$$[1 - h(P)]/2 < \left[1 - h\left(P \| 2^l\right)\right] \leq [1 - h(P)].$$

The approximating partition can be continually improved by approximating higher and higher powers of P. The q^{th} power P^q is the probability distribution obtained by considering sequences of q independent drawings from $\{1, ..., n\}$ with the probabilities P. Thus the probability of a particular sequence $i_1, ..., i_q$ is the product of the probabilities $p_{i_1} ... p_{i_q}$. Since identification probabilities multiply for intersections of independent partitions, $[1 - h(P^q)] = [1 - h(P)]^q$. Each power of P is approximated anew by a complete binary tree as above with the relative block size probabilities for the approximation to P^q denoted by 2^{lq}. Then applying the above inequality, we have:

$$[1 - h(P)]^q/2 = [1 - h(P^q)]/2 < \left[1 - h\left(P^q \| 2^{lq}\right)\right] \leq [1 - h(P^q)] = [1 - h(P)]^q$$

so taking q^{th} roots, we have:

$$\frac{[1-h(P)]}{2^{1/q}} < \left[1 - h\left(P^q \| 2^{lq}\right)\right]^{1/q} \leq [1 - h(P)].$$

Since $2^{1/q} \to 1$ as $q \to \infty$, we have:

$$\lim_{q\to\infty} \left[1 - h\left(P^q \| 2^{lq}\right)\right]^{1/q} = 1 - h(P).$$

The purpose of this section is to show how the setting for Shannon's Noiseless Coding Theorem can be interpreted as a problem about constructing a partition starting at the blob to approximate a given probability distribution. Then the question was treated using the notions of logical entropy rather than the additive Shannon entropy. For comparison purposes, we will now give the Shannon version of the theorem using the same data. We can start with the inequality $\frac{p_i}{2} < \frac{1}{2^{l_i}} \leq p_i$. Taking logs yields $\log(p_i) - \log(2) < -l_i \leq \log(p_i)$ and then taking negatives gives: $\log\left(\frac{1}{p_i}\right) \leq l_i < \log\left(\frac{1}{p_i}\right) + 1$. Then multiplying through by p_i and summing yields:

$$H(P) = \sum_i p_i \log\left(\frac{1}{p_i}\right) \leq \sum_i p_i l_i < \sum_i p_i \log\left(\frac{1}{p_i}\right) + 1 + H(P) + 1$$

where $L = \sum_i p_i l_i$ is the average code word length. The level numbers l_i in the binary tree are now seen as the length of the code word assigned to a node in the code tree (e.g., the code words 0 and 1 are assigned to the two level 1 nodes, $00, 01, 10, 11$ to the four level two nodes and so forth). The Shannon entropy is additive over independent probability distributions so $H(P^q) = qH(P)$ and thus applying the inequality and taking L_q as the average word length in the (Shannon) code for P^q, we have: $qH(P) \leq L_q < qH(P) + 1$ so that dividing through by q yields:

$$H(P) \leq \frac{L_q}{q} < H(P) + \frac{1}{q}.$$

Then taking the limit as $q \to \infty$ yields:

$$\lim_{q\to\infty} \frac{L_q}{q} = H(P).$$

For comparison purposes, it might also be useful to give the theorem for the Shannon multiplicative entropy. Starting again with the inequality $\frac{p_i}{2} < \frac{1}{2^{l_i}} \leq p_i$ we take reciprocals to get $\frac{1}{p_i} \leq 2^{l_i} < \frac{2}{p_i}$ and then raise everything to the power p_i and take the product of corresponding terms to yield:

$$H_m(P) = \prod_i \left(\frac{1}{p_i}\right)^{p_i} \leq \prod_i 2^{p_i l_i} = 2^{\sum p_i l_i} = 2^L < H_m(P) \prod_i 2^{p_i} = H_m(P)\, 2.$$

The multiplicative Shannon entropy is multiplicative for independent probability distributions so $H_m(P^q) = H_m(P)^q$. Thus applying the inequality to P^q yields: $H_m(P)^q \leq 2^{L_q} < H_m(P)^q\, 2$ so taking q^{th} roots yields:

$$H_m(P) \leq \left(2^{L_q}\right)^{1/q} < H_m(P)\, 2^{1/q}.$$

Since $2^{1/q} \to 1$ as $q \to \infty$ we have:

$$\lim_{q\to\infty} \left(2^{L_q}\right)^{1/q} = H_m(P).$$

The three versions of the theorem have essentially the "same" mathematical content; it is a question of which entropy concept is appropriate to the context and interpretation.

4 Concluding Remarks

In the duality of subsets of a set to partitions on a set, we found that the elements of a subset were dual to the distinctions (dits) of a partition. Just as the finite probability theory for events started by taking the size of a subset ("event") S normalized to the size of the finite universe U as the probability $Prob(S) = \frac{|S|}{|U|}$, so it would be natural to consider the corresponding theory that would associate with a partition π on a finite U, the size $|\mathrm{dit}(\pi)|$ of the set of distinctions of the partition normalized by the total number of ordered pairs $|U \times U|$. This number $h(\pi) = \frac{|\mathrm{dit}(\pi)|}{|U\times U|}$ was called the logical entropy of π and could be interpreted as the probability that a randomly picked pair of elements from U (with replacement) is distinguished by the partition π just as $\frac{|S|}{|U|}$ is the probability that a randomly picked element from U is an element of the subset S. Hence this notion of logical entropy arises naturally out of the logic of partitions (a predicate modelled by a partition applies to a dit if the partition makes that distinction) that is dual to the usual logic of subsets (a predicate modelled by a subset applies to an element if the subset contains the element).

The question immediately arises of the relationship with Shannon's concept of entropy. Following Shannon's definition of entropy, there has been a veritable plethora of suggested alternative entropy concepts [8]. Logical entropy is *not* an alternative entropy concept intended to displace Shannon's concept any more than is the multiplicative entropy concept. Instead, I have argued that the logical, multiplicative, and additive concepts of entropy should be seen as three ways to measure that same "information" expressed in its most atomic terms as distinctions. The multiplicative entropy, although it can be independently defined, is trivially related to Shannon's (additive) concept—just take antilogs. The relationship of the logical concept of entropy to the Shannon concept is a little more subtle but is quite simple at the level of blocks $B \in \pi$: $h(B) = 1 - p_B$, $H_m(B) = \frac{1}{p_B}$, and $H(B) = \log\left(\frac{1}{p_B}\right)$ so that eliminating the probability, we have:

$$h(B) = 1 - \frac{1}{H_m(B)} = 1 - \frac{1}{2^{H(B)}}.$$

Then the logical and additive entropies for the whole partition are obtained by taking the (additive) expectation of the block entropies while the multiplicative entropy is the multiplicative average of the block entropies:

$$h(\pi) = \sum_B p_B(1-p_b),\ H(\pi) = \sum_B p_B \log\left(\frac{1}{p_B}\right),\ H_m(\pi) = \prod_B \left(\frac{1}{p_B}\right)^{p_B}.$$

In conclusion, the simple root of the matter is three different ways to "measure" an n-element set. Consider a 4 element set (which can be taken as the discrete partition on that set). One measure of that set is its cardinality 4 and that measure leads to the multiplicative entropy. Another measure of that set is $\log_2(4) = 2$ which can be interpreted as the minimal number of binary partitions necessary: (1) to single out each element as a singleton or, equivalently in the second interpretation, (2) to distinguish all the elements from each other. And the third measure is the (normalized) number of distinctions (counted as ordered pairs) necessary to distinguish all the elements from each other, i.e., $\frac{4\times4-4}{4\times4} = \frac{12}{16} = \frac{3}{4}$. These measures stand in the block value relationship: $\frac{3}{4} = 1 - \frac{1}{4} = 1 - \frac{1}{2^2}$.[18] It is just a matter of counting the distinctions (logical entropy), counting the elements distinguished (multiplicative entropy), or counting the binary partitions needed to distinguish the elements (Shannon entropy).

References

[1] Abramson, Norman 1963. *Information Theory and Coding*. New York: McGraw-Hill.

[2] Baclawski, Kenneth and Gian-Carlo Rota 1979. *An Introduction to Probability and Random Processes*. Unpublished typescript. 467 pages. Download available at: http://www.ellerman.org.

[3] Blackwell, David 1961. Information Theory. In *Modern Mathematics for the Engineer (second series)*. Edwin F. Beckenbach ed., New York: McGraw-Hill: 182-193.

[4] Cover, Thomas and Joy Thomas 1991. *Elements of Information Theory*. New York: John Wiley.

[5] Ellerman, David 2006 A Theory of Adjoint Functors—with some Thoughts on their Philosophical Significance. *What is Category Theory?* Edited by G. Sica. Polimetrica. Milan, 127-183.

[6] Gray, Robert M. 1990. *Entropy and Information Theory*. New York: Springer-Verlag.

[7] Hartley, Ralph V. L. 1928. Transmission of information. *Bell System Technical Journal*. 7 (3, July): 535-63.

[8] Kapur, J.N. 1994. *Measures of Information and Their Applications*. New Dehli: Wiley Eastern.

[9] Lawvere, F. William and Stephen Schanuel 1997. *Conceptual Mathematics: A first introduction to categories*. New York: Cambridge University Press.

[10] Lawvere, F. William and Robert Rosebrugh 2003. *Sets for Mathematics*. Cambridge: Cambridge University Press.

[11] Mac Lane, Saunders 1971. *Categories for the Working Mathematician*. Verlag, New York.

[12] Mac Lane, Saunders and Ieke Moerdijk 1992. *Sheaves in Geometry and Logic: A First Introduction to Topos Theory*. Springer, New York.

[13] Rényi, Alfréd 1965. On the Theory of Random Search. *Bull. Am. Math. Soc.* 71: 809-28.

[14] Rényi, Alfréd 1970. *Probability Theory*. Laszlo Vekerdi (trans.), Amsterdam: North-Holland.

[18] Moreover, it might be noted that these block entropies that stand behind the three concepts make no mention of probabilities. Probabilties enter into the additive and multiplicative Shannon entropies only in the averaging process. Probabilties do not enter into the logical entropy at all (the normalized dit count is only a combinatorial notion) although probabilities are of course involved in the probabilistic interpretation (distinction probability of a random pair).

[15] Rényi, Alfréd 1976. *Selected Papers of Alfréd Rényi: Volumes 1,2, and 3*. Pal Turan (editor), Budapest: Akademiai Kiado.

[16] Schneider, Thomas D. 2005. *Information Theory Primer*, National Cancer Institute. Accessed at: http://www.lecb.ncifcrf.gov/~toms/paper/primer/, October 2006.

[17] Shannon, Claude E. 1948. A Mathematical Theory of Communication. *Bell System Technical Journal*. 27: 379-423; 623-56.

[18] Simpson, Edward Hugh 1949. Measurement of Diversity. *Nature*. 163: 688.

[19] Wood, Richard J. 2004. Ordered Sets via Adjunctions. In *Categorical Foundations. Encyclopedia of Mathematics and Its Applications* Vol. 97. Maria Cristina Pedicchio and Walter Tholen ed., Cambridge: Cambridge University Press: 5-47.